



Advancing drug-response prediction using multi-modal and -omics machine learning integration (MOMLIN): a case study on breast cancer clinical data

Md Mamunur Rashid ¹ and Kumar Selvarajoo ^{1,2,3,*}

¹Biomolecular Sequence to Function Division, BII, (A*STAR), Singapore 138671, Republic of Singapore

²Synthetic Biology Translational Research Program, Yong Loo Lin School of Medicine, NUS, Singapore 117456, Republic of Singapore

³School of Biological Sciences, Nanyang Technological University (NTU), Singapore 639798, Republic of Singapore

*Corresponding author. Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), Singapore 138671, Republic of Singapore; Synthetic Biology for Clinical and Technological Innovation (SynCTI), National University of Singapore (NUS), Singapore 117456, Republic of Singapore; School of Biological Sciences, Nanyang Technological University (NTU), Singapore 639798, Republic of Singapore. E-mail: kumar_selvarajoo@bii.a-star.edu.sg

Abstract

The inherent heterogeneity of cancer contributes to highly variable responses to any anticancer treatments. This underscores the need to first identify precise biomarkers through complex multi-omics datasets that are now available. Although much research has focused on this aspect, identifying biomarkers associated with distinct drug responders still remains a major challenge. Here, we develop MOMLIN, a multi-modal and -omics machine learning integration framework, to enhance drug-response prediction. MOMLIN jointly utilizes sparse correlation algorithms and class-specific feature selection algorithms, which identifies multi-modal and -omics-associated interpretable components. MOMLIN was applied to 147 patients' breast cancer datasets (clinical, mutation, gene expression, tumor microenvironment cells and molecular pathways) to analyze drug-response class predictions for non-responders and variable responders. Notably, MOMLIN achieves an average AUC of 0.989, which is at least 10% greater when compared with current state-of-the-art (data integration analysis for biomarker discovery using latent components, multi-omics factor analysis, sparse canonical correlation analysis). Moreover, MOMLIN not only detects known individual biomarkers such as genes at mutation/expression level, most importantly, it correlates multi-modal and -omics network biomarkers for each response class. For example, an interaction between ER-negative-HMCN1-COL5A1 mutations-FBXO2-CSF3R expression-CD8 emerge as a multimodal biomarker for responders, potentially affecting antimicrobial peptides and FLT3 signaling pathways. In contrast, for resistance cases, a distinct combination of lymph node-TP53 mutation-PON3-ENSG00000261116 lncRNA expression-HLA-E-T-cell exclusions emerged as multimodal biomarkers, possibly impacting neurotransmitter release cycle pathway. MOMLIN, therefore, is expected advance precision medicine, such as to detect context-specific multi-omics network biomarkers and better predict drug-response classifications.

Keywords: multi-omics integration; drug-response prediction; sparse correlation analysis; breast cancer; biomarker discovery

Introduction

The advent of high-throughput sequencing technologies has revolutionized our ability to collect various 'omics' data types, such as deoxyribonucleic acid (DNA) methylations, ribonucleic acid (RNA) expressions, proteomics, metabolomics and bioimaging datasets, from the same samples or patients with unprecedented details [1]. By far, most studies have performed single omics analytics, which capture only a fraction of biological complexity. The integration of these multiple omics datasets offers a more comprehensive understanding of the underlying complex biological processes than single-omic analyses, particularly in human diseases like cancer and cardiovascular disease, where it significantly enhances prediction of clinical outcomes [2, 3].

Cancer is a highly complex and deadly disease if left unchecked, and its heterogeneity poses significant challenges for treatment [4]. Standard treatments, including chemotherapy with or without targeted therapies, aim to reduce tumor burden and improve patient outcomes such as survival rate

and quality of life [5–7]. However, even for the most advanced therapies, such as immunotherapies, treatment effectiveness varies widely across cancer types and even between patients with same diagnosis [8]. This heterogeneity is believed to be due to tumor microenvironment heterogeneity and their effects on the resultant complex and myriad molecular interactions within cells and tissues [9, 10]. This variability underscores the urgent need to identify precise biomarkers to predict individual patient responses and potential adverse reactions to a particular therapy [11]. This can be made possible through multi-omics data integration analyses at the individual patient scale [12].

To assess treatment response, such as pathologic complete response (pCR) and residual cancer burden (RCB), current clinical practice relies on clinical parameters (e.g. tumor size/volume and hormone receptor status), along with genetic biomarkers (e.g. TP53 mutations) [13–15]. However, these approaches do not fully capture the complex intracellular regulatory dynamics [16, 17] or the tumor-immune microenvironment (TIME) interactions

Received: March 25, 2024. Revised: May 30, 2024. Accepted: June 11, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

that influence outcomes [18, 19]. Thus, to enhance personalized cancer treatments, we need novel methodologies that can handle large, complex molecular (omics) and clinical datasets. Machine learning (ML) methods integrating multi-omics data offer a promising avenue to improve prediction accuracy and uncover robust biomarkers across drug-response classes [20], which may be overlooked by single-omics analytics. This approach can predict patients benefiting from standard treatments and those requiring alternative plans like combination therapies or clinical trials.

The current drug-response prediction methods can be broadly categorized into ML-based and network-based approaches. ML methods often analyze each data type (e.g. mutations and gene expression) independently using univariable selection [21, 22] or dimension reduction methods [23]. These results are then integrated using various classifiers or regressors [e.g. support vector machine, elastic-net regressor, logistic regression (LR) and random forest (RF)] [24–26] and ensemble classifier to make predictions [9]. However, these methods often overlooked the crucial interactions among different data modalities. Deep learning methods, while gaining popularity, are limited by the need for large clinical sample sizes to achieve sufficient accuracy [27]. Recent ML advancements have focused on integrating multimodal omics features with patient phenotypes to improve predictive performance [28, 29]. To discover multimodal biomarker, techniques such as multi-omics factor analysis (MOFA) and sparse canonical correlation analysis (SCCA), including its variant multiset SCCA (SMCCA) offer realistic strategies for integrating diverse data modalities [30–32]. However, although these methods are suitable for classification tasks, they are unsupervised and do not directly incorporate phenotypic information (e.g. disease status) to integrate diverse data types. As a result, they are limited to identify phenotype-specific biomarkers.

Recently, advanced supervised approaches like data integration analysis for biomarker discovery using latent components (DIABLO) by Sing et al. (2019) have emerged to overcome these limitations [28]. DIABLO is an extension of generalized SCCA (GSCCA), considers cross-modality relationships and extracts a set of common factors associated with different response categories. Network-based methods, like unsupervised network fusion or random walk with restart approaches construct drug-target interaction and sample similarity networks that are effective for patient stratification [20, 33]. However, these methods lack a specific feature selection design, limiting their utility for identifying biomarkers for patient classification. Nevertheless, none of these ML methods are rigorous in terms of task/class-specific biomarker discovery and interpretability, and both SMCCA and GSCCA struggle with gradient dominance problem due to naive data fusion strategies [34]. Therefore, it is essential to develop novel interpretable methods for identifying robust multimodal network biomarkers across diverse data types to advance our understanding of the complex factors that influence drug responses.

In this study, we introduce MOMLIN, a multi-modal and -omics ML integration framework to enhance the prediction of anticancer drug responses. MOMLIN integrates weighted multi-class SCCA (WMSCCA) that identifies interpretable components and enables effective feature selection across multi-modal and -omics datasets. Our method contributes in three keyways: (i) innovates a class-specific feature selection strategy with SCCA methods for associating multimodal biomarkers, (ii) includes an adaptive weighting scheme into multiple pairwise SCCA models to balance the influence of different data modalities,

preventing dominance during training process and (iii) ensures robust feature selection by employing a combined constraint mechanism that integrate lasso and GraphNet constraints to select both the individual features and subset of co-expressed features, thereby preventing overfitting to high-dimensional data.

We applied MOMLIN to a multimodal breast cancer (BC) dataset of 147 patients comprising clinical features, DNA mutation, RNA expression, tumor microenvironment and molecular pathway data [9], to predict drug-response classes, specifically distinguishing responders and non-responders. Our results demonstrate MOMLIN's superiority in terms of outperforming state-of-the-art methods and interpretability of the underlying biological mechanisms driving these distinct response classes.

Background and methods

Overview of our proposed method for treatment response prediction

The workflow of our proposed method MOMLIN for identifying class- or task-specific biomarkers from multimodal data is shown in Fig. 1. The core of this pipeline involves three stages: (i) identification of response-specific sparse components, in terms of input features and patients, (ii) development of drug-response predictor using latent components of patients and (iii) interpretation of sparse components and multi-modal and -omics biomarker discovery.

The rationales underpinned of this approach is that effective biomarkers are: (i) response-related multimodal features including genes, cell types and pathways, and (ii) features that demonstrate prediction capabilities on unseen patients. The first stage, a 'feature selection step' that selects multimodal features on the generated sparse components based on their relevance to drug-response categories (pCR and RCB-I to III). Features with high loading identified are considered as potential biomarker candidates. The second stage, a 'classification step', validates these biomarkers by assessing their predictive power in distinguishing responders from non-responders to anticancer therapy; any predictions indicating chemo-resistant tumors should be considered for enrolment in clinical trials for novel therapies. The third stage, an 'interpretation step,' analyzes the candidate biomarkers in a multi-modal and -omics network associated with relevant biological pathways. This step aims to elucidate the underlying biological processes differentiating between drug-response phenotypes.

Stage 1. Identification of response-associated sparse components in terms of input features and patients

Multi-modal and -omics data overview and preparation

This study utilized clinical attributes, DNA mutation and gene expression (transcriptome) data from 147 matched samples of early and locally advanced BC patients (categorized as pCR, $n = 38$, RCB-I, $n = 23$, or RCB-II, $n = 61$, or RCB-III, $n = 25$), obtained from the TransNEO cohort at Cambridge University Hospitals NHS Foundation [9]. The dataset includes clinical attributes (8 features, summary attributes are available in [Supplementary Table S1](#) available online at <http://bib.oxfordjournals.org/>), genomic features (31 DNA mutation genes, applying a strict criterion of genes mutated in at least 10 patients) and RNA-sequencing (RNA-Seq) features (18 393 genes), covering major BC subtypes-normal-like, basal-like, Her2, luminalA and luminalB. Although DNA mutation genes typically represent binary data, we used mutation frequencies to construct a mutation count matrix. Initial data

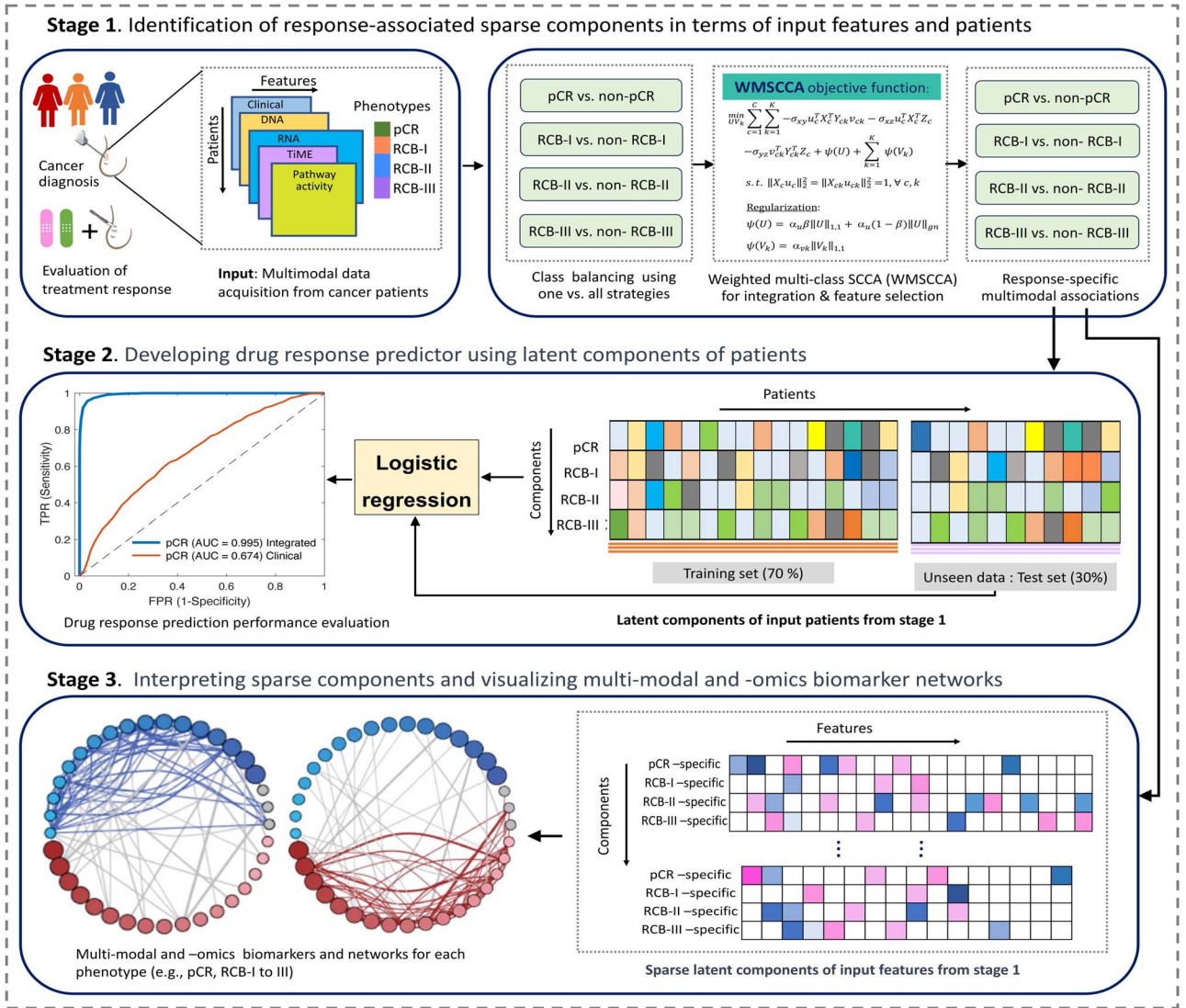


Figure 1. Schematic representation of the proposed framework. In stage 1, multimodal datasets from cancer patients (e.g. BC) were sourced from a published study [9]. This dataset comprises clinical features, DNA mutations, and gene expression from pre-treatment tumors, alongside post-treatment response classes (pCR, RCB-I to III). TIME and pathway activity were derived from transcriptomic data using statistical algorithms. For identifying class-specific correlated biomarkers, class binarization and oversampling were used to balance between classes. WMSCCA models the multimodal associations across different biomarkers and identifies response-specific sparse components on diverse input features and patients. In stage 2, a binary LR classifier then utilizes these patient latent components for predicting response to therapies, evaluated by AUROC. Next in stage 3, class-specific sparse components are shown in a heatmap, highlighting key signatures (non-zero loading) in colors. Finally, the identified multi-modal and -omics signatures then formed a correlation network, revealing pathways associations with multi-modal and -omics biomarkers for each response class. Nodes with colors in the network indicate multimodal features.

pre-processing involved a log₂ transformation on the RNA-Seq features after filtering out less informative features at 25th percentile (in terms of mean and standard deviation) using interquartile range. For integrative modeling, we used the top 40% of variable genes (3748 genes, based on median absolute deviation ranking) from the RNA-Seq datasets. Finally, each feature was normalized dividing by its Frobenius norm, adjusting the offset between high and low intensities across different data modalities.

To characterize TIME and pathway markers, we applied various statistical algorithms on the RNA-Seq data. The GSVA algorithm [35] calculated (i) the GGI gene sets [36] and (ii) STAT1 immune signature scores [37]. For immune cell enrichment, three methods were used: (i) MCPcounter [37] with voom-normalized RNA-Seq counts; (ii) enrichment over 14 cell types using 60 gene markers, employing log₂-transformed geometric mean of transcript per million (TPM) expression [38]; and (iii)

z-score scaling of cancer immunity parameters [39] to classify four immune processes (major histocompatibility complex molecules, immunomodulators, effector cells and suppressor cells). Additionally, the TIDE algorithm [40] computed T-cell dysfunction and exclusion metrics for each tumor sample using log₂-transformed TPM matrix of counts, which can serve as a surrogate biomarker to predict the response to immune checkpoint blockade. Pathway activity scores for each tumor sample were computed using the GSVA algorithm with input gene sets from Reactome [41], PIP [42] and BioCarta databases within the MSigDB C2 pathway database [43].

Sparse multiset canonical correlation analysis

In this study, lowercase letters denote a vector, and uppercase ones denote matrices, respectively. The term $\|\cdot\|_{1,1}$ denotes the matrix l_1 -norm, and $\|\cdot\|_{gn}$ denotes the GraphNet regularization.

The sparse multiset canonical correlation analysis (SMCCA) is an extension of dual-view SCCA, proposed to model associations among multiple types of datasets [31]. Given the multiple types of datasets, let $X \in \mathcal{R}^{n \times p}$ represent gene expression data with p features, and $Y_k \in \mathcal{R}^{n \times q_k}$ represent the k -th data modality (e.g. clinical, DNA mutation and tumors microenvironment) with q_k features. Both X and Y_k have n samples, and $k = (1, \dots, K)$, where K denotes the number of different data modalities. The objective function of SMCCA is defined as follows:

$$\min_{u, v_k} \sum_{k=1}^K -u^T X^T Y_k v_k + \lambda_u \|u\|_1 + \sum_{k=1}^K \lambda_{v_k} \|v_k\|_1 \quad (1)$$

$$\text{s.t. } \|u\|_2^2 = 1, \|v_k\|_2^2 = 1, \forall k,$$

where u and v_k are the canonical weight vectors corresponding to X and Y_k , indicating the importance of each respective biomarkers. The term $\| \cdot \|_1$ represents the l_1 regularization to detect small subset of discriminative biomarkers and prevent model overfitting. λ_u, λ_{v_k} are non-negative tuning parameters balancing between the loss function and regularization terms. The term $\| \cdot \|_2^2$ denotes the squared Euclidean norm to constraint weight vectors u and as unit length v_k , respectively.

However, SMCCA has limitations: (i) it is naturally unsupervised, meaning SMCCA cannot leverage phenotypic information (e.g. disease status and drug-response classes); (ii) pairwise association among multiple data types can vary significantly and can lead to gradient dominance issues during optimization; and (iii) SMCCA mines a common subset of biomarkers for classifying different tasks, which diminishes its relevance, as each task might require distinct features sets.

Weighted multi-class sparse canonical correlation analysis

To address the above limitations, here we propose weighted multi-class SCCA (WMSCCA), a formal model for class/tasks-specific feature selection, different from the conventional SMCCA. Throughout this study, we used the terms tasks/classes/drug-response classes interchangeably. WMSCCA includes phenotypic information as an additional data type, employs a weighting scheme to resolve the gradient dominance issue and innovates traditional class-specific feature selection strategies through the one-versus-all strategies into its core objective function. In this study, the underlying motivation is WMSCCA can jointly identify drug-response class-specific multimodal biomarkers to improve drug-response prediction. For ease of presentation, we consider n patients with data matrices $X_c \in \mathcal{R}^{n \times p}$, $Y_{ck} \in \mathcal{R}^{n \times q_k}$, and $Z \in \mathcal{R}^{n \times c}$ from C different drug-response classes. Here, X_c denotes p features from gene expression datasets, Y_{ck} denotes q_k features from k -th data modality (e.g. mutation, clinical features, TIME and pathway activity), Z_c denotes c response class, and $k = (1, \dots, K)$, K denotes the number of data modalities. The WMSCCA optimization problem can be formulated as follows:

$$\min_{U, V_k} \sum_{c=1}^C \sum_{k=1}^K -\sigma_{xy} u_c^T X_c^T Y_{ck} v_{ck} - \sigma_{xz} u_c^T X_c^T Z_c - \sigma_{yz} v_{ck}^T Y_{ck}^T Z_c \quad (2)$$

$$+ \psi(U) + \sum_{k=1}^K \psi(V_k), \quad (2)$$

$$\text{s.t. } \|X_c u_c\|_2^2 = 1, \|Y_{ck} v_{ck}\|_2^2 = 1, \text{ for all } c, k.$$

where $U \in \mathcal{R}^{p \times C}$, $V_k \in \mathcal{R}^{q_k \times C}$ are canonical loading matrices correspond to X and Y_k , representing the importance of candidate

biomarkers for each class C , respectively. In this equation, the first term models associations among X , and Y_k datasets; the second- and third terms correlate class labels Z_c with X and Y_k data modalities for each C^{th} class, aiming to identify class-specific features and their relationships; $\psi(U)$ and $\psi(V_k)$ represent sparsity constraints on U and V_k , to select a subset of discriminative feature. As mentioned in Equation (1), to address gradient dominance, the adjusting weight parameter σ_{xy} , σ_{xz} and σ_{yz} can be defined as:

$$\sigma_{xy} = \frac{1}{\|Xu - Y_k v_k\|_2}, \sigma_{xz} = \frac{1}{\|Xu - Z\|_2}, \sigma_{yz} = \frac{1}{\|Y_k v_k - Z\|_2}, \quad (3)$$

where $k = (1, \dots, K)$, K denotes the number of data modalities. σ_{\cdot} adjusts a larger weight if the non-squared loss (denominator term) between datasets is small and vice versa.

Given high-dimensional datasets, the model in Equation (2) encounters an overfitting problem. Therefore, the use of a sparsity constraint is appropriate to address this issue. We hypothesized that gene expression biomarkers can be either single genes or co-expressed sets; thus, a combined penalty is designed for the X dataset. Therefore, $\psi(U)$ for X takes the following form:

$$\psi(U) = \alpha_u \beta \|U\|_{1,1} + \alpha_u (1 - \beta) \|U\|_{gn}, \quad (4)$$

where, α_u, β are nonnegative tuning parameters. β balances between the effect of co-expressed and individual feature selection. The first sparsity constraint is matrix $l_{1,1}$ -norm, which is defined as follows:

$$\|U\|_{1,1} = \sum_{i=1}^p \sum_{c=1}^C |u_{ic}| \quad (5)$$

This penalty promotes class-specific features on U . The second sparsity constraint GraphNet regularization, defined as follows:

$$\|U\|_{gn} = \sum_c u_c^T L_c u_c, \quad (6)$$

where L_c represents the Laplacian matrices of the connectivity in X matrices. The Laplacian matrix is defined as $L = D - A$, where D is the degree matrix of connectivity matrix A (e.g. gene co-expression or correlation network). This penalty term promotes a subset of connected features to discriminate each response on U .

Besides, neither every mutation marker nor every clinical/TIME/pathways involves in predicting response classes, therefore, the $l_{1,1}$ -norm is used on the Y_k datasets to select individual markers, i.e. $\psi(V_k)$ for the Y_k data modalities take the following form:

$$\psi(V_k) = \alpha_{v_k} \|V_k\|_{1,1}, \quad (7)$$

where α_{v_k} is non-negative tuning parameter.

Finally, we obtained C pairs of canonical weight matrices ($U_c V_{ck}$) ($c = 1, \dots, C$; $k = 1, \dots, K$) using an iterative alternative algorithm by solving Equation (2) [44, 45]. Detected features with non-zero weights in each class in the weight vectors were extracted as correlated sets.

The WMSCCA method involves parameters α_u, β , and α_{v_k} ($k = 1, \dots, K$). Given the limited number of samples, we applied a nested cross-validation (CV) strategy on training sets and evaluated the maximum correlation on the test datasets. Optimal

values for the regularization parameters were determined within each training set via internal five-fold CV.

Stage 2. Drug-response prediction using latent components of patients

To predict drug-response categories, we trained LR classifier using the latent components of patients (or raw multimodal features) generated by MOMLIN in Fig. 1: stages 1 and 2. We used a binary classification scheme, distinguishing pCR versus non-pCR, RCB-I versus non-RCB-I, RCB-II versus non-RCB-II and RCB-III versus non-RCB-III, to evaluate model performance. In addition, we performed analyses with existing multi-omics methods, including SMCCA+LR, MOFA+LR, DIABLO and latent principal component analysis (PCA) features, with LR classifiers. To assess prediction performance for the response to treatment in an unbiased manner, we used five-fold cross-validated performance and repeated the process over 100 runs. The partitioning of data was kept consistent across all models for fair comparisons. The accuracy of response prediction was evaluated using area under the receiver operating characteristic curve (AUROC).

Stage 3. Interpretation of sparse components and multi-omics biomarker discovery and their networks

After learning sparse latent components of features across different data modalities using MOMLIN, we identify the most relevant feature based on the loading weight of genes, TiME and pathways, which reveal underlying interactions for discriminating response classes. The larger the loading weight, the more important the pair of features in discriminating response categories. We then use these selected features to construct a sample correlation network, or a relationship matrix based on their canonical weights [46]. In this network, nodes represent selected features, and the edge weights between two interconnected features indicate correlation or relatedness. The generated network is visualized using the ggraph package in R (<https://cran.r-project.org>). Finally, we prioritize multi-omics biomarkers based on their degree centrality within the interconnected correlation network.

Results

Derivation of response-associated latent components from BC data with MOMLIN

We applied MOMLIN to analyze a breast cancer (BC) dataset to predict treatment response and gain molecular insights. The dataset comprised 147 BC patients with early and locally advanced pretherapy tumors [9], categorized as follows: pCR with 38 patients, RCB-I (good response) with 23 patients, RCB-II (moderate response) with 61 patients and RCB-III (resistance) with 25 patients. After preprocessing and filtering least informative features, the final dataset comprised 3748 RNA genes (top 40% out of 9371 genes), 31 mutation genes, 8 clinical attributes, 64 TiME and 178 pathways activities (Fig. 1: stage 1). [Supplementary Table S1](http://bib.oxfordjournals.org/) available online at <http://bib.oxfordjournals.org/> summarizes overall clinical characteristics by patients' response classes.

While our proposed framework offers general applicability for identifying context-specific multi-omics biomarkers, this study specifically focused on discovering drug-response-specific biomarkers to enhance the prediction of pCR and RCB resistance. MOMLIN decomposed the input multimodal data into response-associated sparse latent components of input-features and

patients. These sparse components reveal patterns of how various features (e.g. genes and mutations) and clinical attributes related to treatment outcomes (Fig. 1: stage 1–3), and their effectiveness was evaluated by measuring prediction performance. We assessed the predictive ability of MOMLIN through five-fold CV repeated 100 times. In each iteration, the dataset is divided into five-folds, with one random fold assigned as the held-out test set, and the remaining folds used as the training set. MOMLIN was trained using the training dataset, including detection of predictive marker candidates, and its performance was evaluated on the 'unseen' test set. This process was repeated for all five-folds to ensure robust evaluation of MOMLIN's generalizability. Performance was measured by the AUROC matrices (Fig. 1: stage 2).

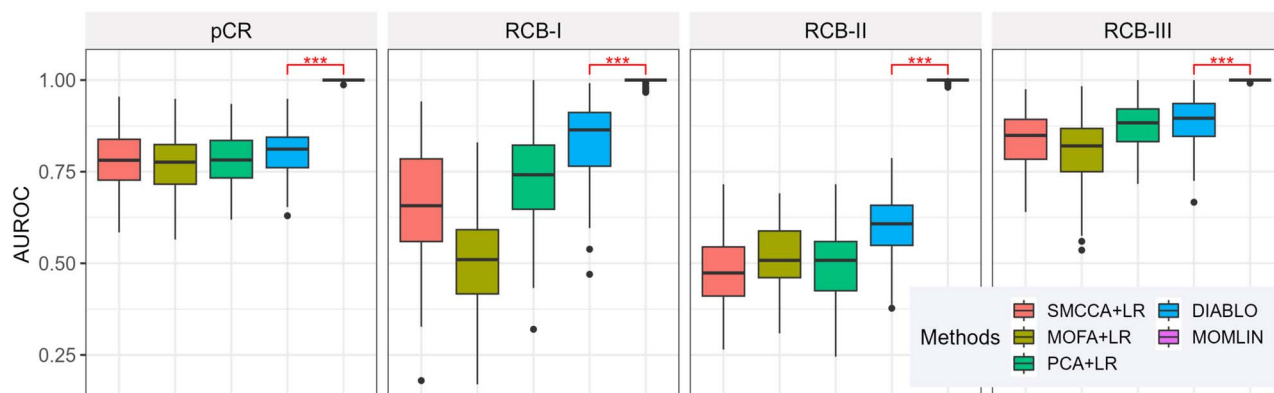
Performance comparison with existing methods for drug-response prediction

To evaluate the prediction capability of MOMLIN, we modeled each response category as a binary classification problem and compared its prediction accuracy to existing multi-omics integration algorithms. For comparison, we randomly split the dataset into a training set (70%) and a test set (30% unseen data), with balanced inclusion of response classes. We employed LR as the classifier to assess predictive performance of multimodal biomarkers. We compared MOMLIN with four other classification algorithms for omics data: (i) SMCCA, which integrates multi-omics data by projecting it onto latent components for discriminant analysis; (ii) MOFA, which decomposes multi-omics data into common factors for discriminant analysis; (iii) sparse PCA; and (iv) DIABLO, a supervised integrative analysis method, represent the state-of-the-art in classification. All methods were trained on the same preprocessed data.

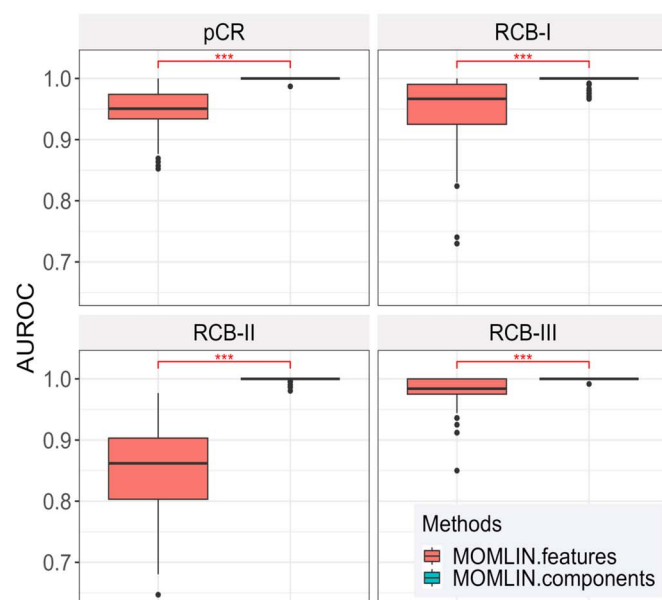
The classification results showed that MOMLIN outperformed the compared multi-omics integration methods in most classification tasks on unseen test samples (Fig. 2A). Notably, DIABLO, the next best performer, was 10 to 15% less effective than our MOMLIN. Additionally, we compared the performance of component-based LR models against raw feature-based LR models to predict RCB response classes. Although raw feature-based models showed improved prediction, their performance was notably dropped compared to component-based models (Fig. 2B). This indicates the superior adaptability and effectiveness of component-based models in leveraging multi-omics data for predictive purposes.

Moreover, to test and demonstrate generalizability of this framework, we applied MOMLIN to a preprocessed multi-omics dataset of colorectal adenocarcinoma (COAD) with 256 patients [47]. This dataset included gene expression, copy number variations and micro-RNA expression data, which we used to classify COAD subtypes such as chromosomal instability (CIN, $n=174$), genomically stable (GS, $n=34$) and microsatellite instability (MSI, $n=48$). The performance results shown in [Supplementary Table S2](http://bib.oxfordjournals.org/) available online at <http://bib.oxfordjournals.org/> and [Supplementary Figure S1](http://bib.oxfordjournals.org/) available online at <http://bib.oxfordjournals.org/>, indicate that MOMLIN outperformed all state-of-the-art methods tested in classifying COAD subtypes. Moreover, when comparing the raw feature-based accuracies with sparse components-based (features derived from MOMLIN) accuracies, we found that raw feature-based classifier was superior against existing methods (Figure S1A and B), but lower than the components-based classifier. This consistent observation supports our findings with BC drug-response performances.

A. Different methods performance comparisons on unseen data (test data)



B. Performance comparison based on raw features and detected sparse components



C. Informative modality detection by data integration

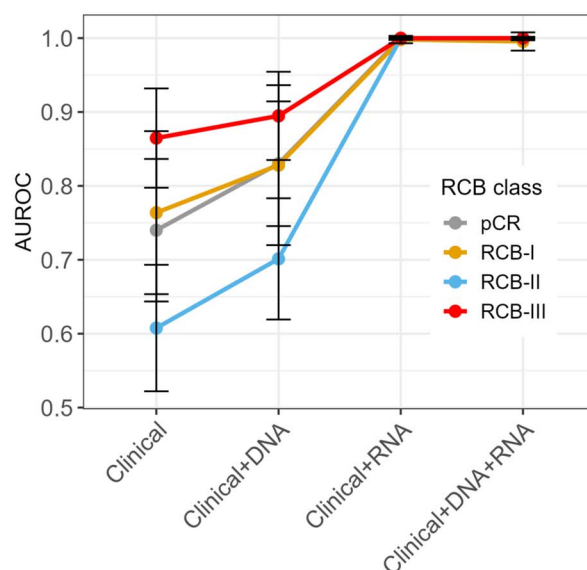


Figure 2. Performance comparison with existing methods and detection of informative data combination. All results in the plots depict test AUROC over five-fold CV obtained from 100 runs. (A) Box plots comparing response prediction performance of MOMLIN against existing state-of-the-art multi-omics methods. (B) Performance comparison between predictors based on latent components and those utilizing a selected subset of multimodal features. (C) Comparing AUROCs for the models with different data subset combinations (clinical, clinical + DNA, clinical + RNA and clinical + DNA + RNA) using MOMLIN.

Importance of different omics data for treatment response prediction

To assess the added value of integrating multimodal data for predicting treatment response, we trained four prediction models with different feature combinations: (i) clinical features only, plus adding (ii) DNA, (iii) RNA and (iv) both DNA and RNA. We found that adding different data modalities improved prediction performance across all response classes (Fig. 2C). Notably, the models that combined clinical data with either RNA or both DNA and RNA demonstrated superior and comparable performance with an average AUROC of 0.978. In contrast, the model based on clinical features alone had much lower AUROC, ranging from 0.51 to 0.82. These results suggest that RNA transcriptome is the most informative data modality in this dataset. Thus, integrating gene expression with clinical features could significantly improve our ability to predict treatment outcomes in BC.

Interpretation of response-associated sparse components identified by MOMLIN

To understand the molecular landscape of treatment response in BC, we used MOMLIN to model response-specific bi-multivariate associations across multiple data modalities. We observed stronger correlations between RNA gene expression and both TIME ($r=0.701$) and pathway activity ($r=0.868$), indicating greater overlap or explained information between them. Conversely, moderate correlations were found between RNA gene expression and DNA mutations ($r=0.526$), or clinical features ($r=0.488$), indicating partially overlapping or independent information. These results suggest that multimodal biological features provide complementary information in a combinatorial manner.

When investigating the importance of each feature to predict response classes, MOMLIN identified four distinct loading vectors corresponding to pCR and RCB response classes,

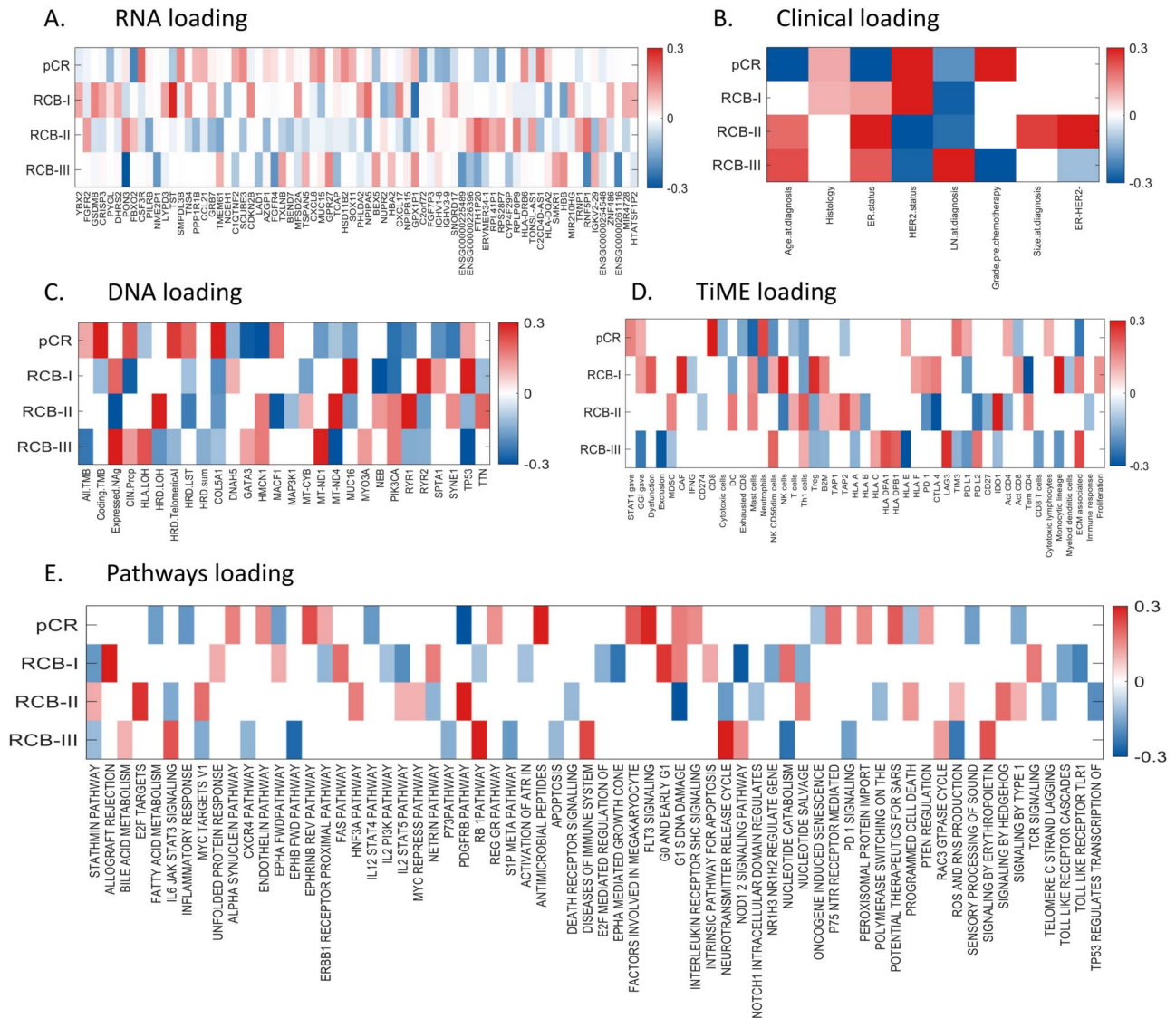


Figure 3. Heatmaps illustrate the features importance on response-associated components identified by MOMLIN. Each row in the heatmap represents a drug-response class, pCR, RCB-I, RCB-II and RCB-III, with columns representing features across different data modalities. The color gradient indicates feature loading or importance, representing the strength of association with response classes. The sign (negative or positive) of gradient denotes the association directions to response classes. All results in the heatmaps depict an average over 100 runs of five-fold CV. (A–E) represents the response-associated candidate biomarkers detected in latent components in (A) gene expression data (highlighting DE genes), (B) clinical features, (C) DNA mutations (highlighting mutated genes), (D) TIME cells and (E) functional pathway profiles (highlighting altered pathways).

highlighting distinct weight patterns for pCR versus non-pCR (and RCB versus non-RCB classes (Fig. 3). For example, in the pCR (complete response) components—taking the top five molecular features across different modalities revealed distinct molecular patterns. Specifically, gene expression analysis showed that downregulation of FBXO2 and RPS28P7 inhibits tumor cell proliferation, and potentially may enhance treatment efficacy, and the upregulation of C2CD4D-AS1, CSF3R, and SMPDL3B genes may promote immune response, increasing tumor cell vulnerability and therapeutic effect (Fig. 3A). Mutational analysis revealed negative associations of marker genes HMCN1 and GATA3, but a positive association for COL5A1 (Fig. 3C). Additionally, tumor mutation burden (TMB), and homologous recombination deficiency (HRD)-Telomeric AI signatures were higher in pCR patients, suggesting high genomic instability compared to RCB patients [9]. TIME analysis showed reduced immunosuppressive mast cells and extracellular matrix (ECM), along with increased

infiltration of neutrophils, TIM-3 and CD8+ T-cells (Fig. 3D). Subsequently, the pathway analysis further revealed potential downregulation of the PDGFRB pathway, involved in stromal cell activity and associated with improved patient response [49], while upregulation of pathways for antimicrobial peptides, FLT3 signaling, ephrin B reverse signaling and potential therapeutics for SARS (Fig. 3E), suggesting enhanced immune surveillance and interaction with tumor cells. In summary, MOMLIN reveals distinct genomic landscape with higher immune activity and genomic instability in pCR that characterizes its favorable treatment response.

Similarly, in the RCB-I (good response) components—RNA expression analysis revealed that lower expression of genes GPX1P1 and HBB are linked to less aggressive tumors [48], while those of thiosulfate sulfurtransferase (TST), NPIPA5 and GSDMB were overexpressed, linked to enhanced immune response and therapeutic effectiveness [49, 50]. Mutational analysis showed

positive association for therapeutic targets signatures TP53, MUC16 and RYR2 [51, 52], but a negative in NEB, and CIN scores. TiME analysis demonstrated increased infiltration of Tregs, cancer-associated fibroblast (CAF), monocytic lineage and natural killer (NK) cells, indicating more active of immune environment [9], with reduced TEM CD4 cells. Pathway analysis further identified downregulation of NOD1/2 signaling, EPHA-mediated growth cone collapse and toll-like receptor (TLR1, TLR2) pathways, involved in inflammation and immune response, with the upregulation of allograft rejection, and G0 and early G1 pathways. In summary, tumors that achieve RCB-I is marked by distinct genomics marker, active immune response, and lower CIN.

In RCB-II (moderate response) components: RNA expression analysis revealed overexpression of RPLP0P9, FTH1P20, RNF5P1 pseudogenes, following accumulation of overexpressed ERVMER34-1, and PON3 genes play an oncogenic role in BC [53]. Mutation analysis revealed positive association of HRD-LOH, RYR1 and MT-ND4, but negative association of MACF1 and neoantigen loads, in line with previous reports [54, 55]. Analysis of TiME features demonstrated increased infiltration of IDO1 and TAP2, with reduced CTLA 4, NK cells and PD-L2 cells, indicating a less suppressive immune environment. Pathways analysis further revealed downregulation pathways of G1/S DNA damage checkpoints and TP53 regulation, highlighting DNA repair issues, with the upregulation of PDGFRB pathway, E2F targets and signaling by Hedgehog associated with cell proliferation. In summary, RCB-II patients display distinct genomics markers including pseudogenes, lack of suppressive immune environment and active proliferation.

In RCB-III (resistant) components: RNA gene expression analysis revealed lower expression of therapeutic target PON3, and FGFR4 [56], and flowed accumulation of lower expressed lncRNAC ENSG00000225489, ENSG00000261116 and RNF5P1. Mutation signature analysis identified a positive association of MT-ND1, but a negative association in therapeutic targets TP53, and MT-ND4 [7, 52]. Neoantigen loads were higher following lower TMB indicate reduced tumor suppressor activity. TiME analysis revealed reduced activity of T-cell exclusion, and HLA-E, with increased ECM, HLA DPA1 and LAG3, suggesting an immune suppressive tumor environment. Pathway analysis revealed upregulation of pathways involved in neurotransmitter release, cell-cycle progression (RB-1) and immune system diseases, suggesting active cell signaling and proliferation, with downregulation of EPHB FWD pathway and nucleotide catabolism. In summary, patients that attained RCB-III, characterized by low mutational burden and an immune suppressive environment, leading to treatment resistance.

Linking biology to treatment response through biomarker network analysis

To further extract multimodal network biomarkers and understand the complex biological interactions in patients with pCR and RCB, we performed cross-interaction network analysis using candidate signatures identified by MOMLIN across different modalities. This analysis included clinical features, DNA mutations, gene expression, TiME cells and enriched pathways, aiming to elucidate the underlying biology associated with specific treatment responses. Figure 4 shows the interaction networks of selected multimodal features for each RCB class. To identify potential biomarkers associated with pCR and RCB response, we specifically focused on the top ten multimodal features based on network edge connections. For example, tumors that attained in

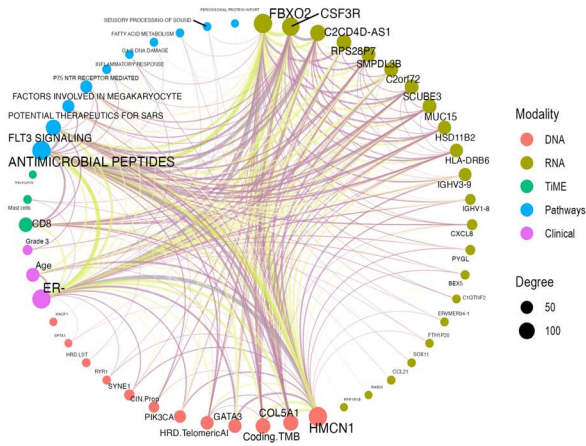
pCR, the network analysis revealed co-enrichment of mutations in HMCN1 and COL5A1 genes, particularly in estrogen receptor (ER)-negative patients. HMCN1 and COL5A1 therapeutic targets like molecules encode proteins for ECM structure, and mutations of these genes regulate tumor architecture and cell adhesion, potentially facilitating immune cell infiltration [52]. We also observed elevated expressions of FBXO2, CSF3R, C2CD4D-AS1 and RPS28P7 genes, alongside increased infiltration of CD8+ T-cells [9, 57]. FBXO2 is a component of the ubiquitin-proteasome system, which regulates protein degradation and influences cell cycle and apoptosis [58], while CSF3R plays a vital role in granulocyte production and immune response [59]. These gene expression patterns, coupled with increased CD8+ T-cell infiltration, suggest a robust anti-tumor immune response. Furthermore, these molecular perturbations may be linked to antimicrobial peptide pathways and FLT3 signaling, potentially contributing to the favorable outcome in achieving pCR [60, 61]. Future work could specifically search for these complex interactions across different molecules to gain more clinically relevant insights into pCR tumors. Supplementary Table S3 available online at <http://bib.oxfordjournals.org/> presents the more detailed list (top 30) of the multi-modal and -omics biomarkers identified using the MOMLIN pipeline.

Similarly, RCB-I tumors exhibited co-enriched mutations in MUC16 and TP53, particularly in HER2+ cases [14]. MUC16 (CA125) is therapeutic molecule associated with immune evasion and tumor growth [51], while TP53 mutations can lead to loss of cell cycle control and genomic instability [62]. We also observed elevated expression of TST involved in the detoxification processes and GPX1P1 [long non-coding RNA (lncRNA)] involved in oxidative stress response. The immune landscape of these tumors showed increased infiltration of TEM CD4 cells (adaptive immunity), monocytic lineage cells (phagocytosis and antigen presentation) and NK cells (innate immunity), as well as CAFs. This immune landscape, coupled with potential perturbations in the allograft rejection pathway, suggests an active but potentially incomplete immune response against the tumor, resulting in minimal residual disease.

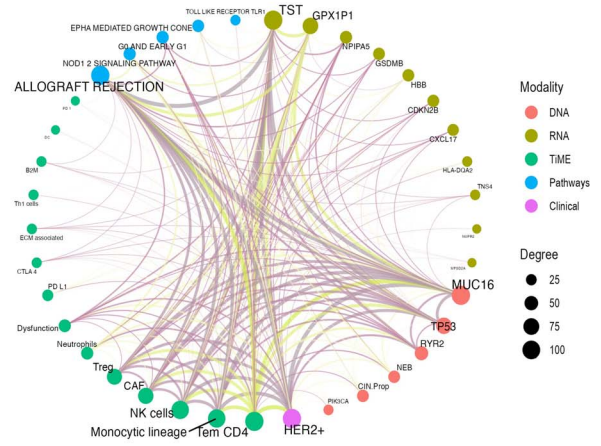
RCB-II tumors had lower neoantigen loads compared to pCR, both in ER-negative and HER2+ patients. This reduced neoantigen load might contribute to a weaker immune response. Gene expression analysis showed elevated levels of specific lncRNAs, including FTH1P20 (associated with iron metabolism), RNF5P1 (potentially affecting protein degradation) and RPLP0P9 (involved in protein synthesis), along with ERVMER34-1, which can influence gene expression and immune response in BC patients. Numerous studies have underscored the key regulatory roles of lncRNAs in tumors and the immune system. Notably, increased expression of the immune checkpoint protein IDO1 negatively regulates the expression of CTLA-4, both known to modulate antitumor immune responses [63]. The combined effect of these molecular alterations suggests potential tumor survival mechanisms, including immune evasion and dysregulation of G1/S DNA damage [64] contributing to moderate residual disease.

In RCB-III tumors, we observed the reduced prevalence of TP53 and MT-ND4 mutations, typically associated with genomic instability and aggressive tumor behavior [51], coupled with a higher neoantigen load, suggesting an alternative mechanism (pathways) that drives tumor progression. Despite the higher neoantigen loads, increased expression of HLA-E immune checkpoints and T-cell exclusion in the tumor microenvironment hindered effective anti-tumor immune responses. Additionally, the low-expressed genes PON3, ENSG00000261116 (lncRNA) and RNF5P1

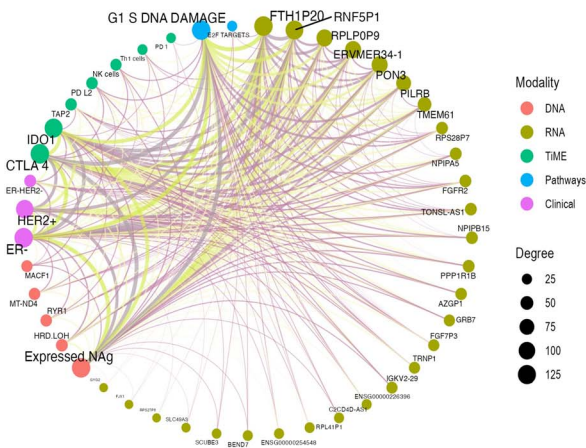
A. pCR biomarkers and networks



B. RCB-I biomarkers and networks



C. RCB-II biomarkers and networks



D. RCB-III biomarkers and networks

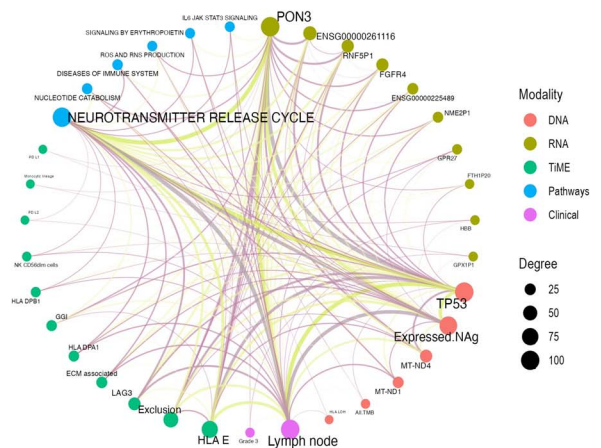


Figure 4. Multimodal network biomarkers explain drug-response classes. The multimodal networks detail the candidate biomarkers and their interactions for each response class, (A) the pCR patients (B) the RCB-I patients (good response), (C) the RCB-II patients (moderate response) and (D) the RCB-III resistance patients. Nodes in the network represent candidate biomarkers derived from clinical features, DNA mutations, gene expression, enriched cell-types and pathways, each indicated in different colors in the figure legend. Negative edges are light green; positive edges are in light magenta. Edge width reflects the strength of the interaction between features. Node size corresponds to the number of connections (degree), and the font size of node labels scales with degree centrality, highlighting the most interconnected biomarkers.

are involved in detoxification, gene regulation and protein degradation, respectively, represents an adaptive response to cellular stress in these tumors. Clinical markers indicating lymph node involvement suggest a more advanced disease state [9]. These findings, along with potential perturbations in the neurotransmitter release cycle pathway, collectively portray RCB-III tumors as genetically unstable, yet effectively evading immune surveillance, contributing to their significant treatment resistance. Overall, further investigation of these interactive molecular networks, comprising both positive and negative interactions offers a more depth understanding of these potential candidate biomarkers for distinguishing treatment-sensitive pCR and resistant RCB tumors.

Discussion

The advent of multi-omics technologies has revolutionized our understanding of cancer biology, offering unprecedented insights into the complex molecular interactions that shape tumor behavior and treatment response. In this study, we presented MOMLIN (multi-modal and -omics ML integration),

a novel method to enhance cancer drug-response prediction by integrating multi-omics data. MOMLIN specifically utilizes class-specific feature learning and sparse correlation algorithms to model multi-omics associations, enables the detection of class-specific multimodal biomarkers from different omics datasets. Applied to a BC multimodal dataset of 147 patients (comprising RNA expression, DNA mutation, tumor microenvironment, clinical features and pathway functional profiles), MOMLIN was highly predictive of responses to anticancer therapies and identified cohesive multi-modal and -omics network biomarkers associated with responder (pCR) and various levels of RCB (RCB-I: good response, RCB-II: moderate response and RCB-III: resistance).

Using MOMLIN, we identified that pCR is determined by an interactive set of multimodal network biomarkers driven by distinct genetic alterations, such as HMCN1 and COL5A1, particularly in ER-negative tumors [9, 65]. Gene expression signatures, including FBXO2 and CSF3R were associated with the immune cell infiltration (CD8+ T-cells), which has been previously reported as a key determinant of response [57]. The association of these biomarkers with antimicrobial peptide and FLT3 signaling pathways suggests a robust immune response [61] as a critical

driver of complete response. Additionally, C2CD4D-AS1, an lncRNA was identified, and its exact role with these complex molecular interactions in BC remains to be elucidated. Future work could specifically search for these complex interactions across different molecules to gain more clinically relevant insights into pCR tumors.

RCB-I tumors, despite responding well to response, were associated with a distinct multimodal molecular signature. These tumors were enriched for mutations in the therapeutic target MUC16 (CA125), known for its role in immune evasion [51], and the tumor suppressor gene TP53, particularly in HER2+ cases [14]. Elevated expression of TST and GPX1P1 (lncRNA involved in oxidative stress response) were associated with increased infiltration of diverse immune cells, including Tem CD4+ cells, monocytes and NK cells [10]. This active immune landscape and the intricate interactions of these signature with the potential perturbations in the allograft rejection pathway, suggests a robust yet potentially incomplete anti-tumor immune response, contributing to the minimal residual disease observed in this subtype.

RCB-II tumors showed lower neoantigen loads compared to pCR, which could contribute to a weaker immune response, particularly in ER-negative and HER2+ subtypes. Increased expression of lncRNAs, such as FTH1P20, RNF5P1, RPLP0P9 and ERVMER34-1, were associated with the immune checkpoint protein IDO1, and negatively regulate the CTLA-4 protein expression, suggests immune evasion and alterations in tumor cell metabolism and proliferation. These molecules altered intricate interactions implicate dysregulation of G1/S DNA damage as a possible mechanism for moderate treatment response [64].

RCB-III tumors, classified as resistant, were associated with a distinct multimodal molecular landscape driven by reduced TP53 and MT-ND4 mutations [52], accompanied with higher neoantigen loads compared to other response groups. This suggests an alternative mechanism driving tumor progression and immune evasion. Despite the high neoantigen load which could potentially trigger immune response, these tumors exhibited immune evasion through increased HLA-E immune checkpoints and T-cell exclusion [40, 55]. Also, the downregulation of genes like PON3 and the lncRNA ENSG00000261116, along with lymph node involvement, pointed to advanced disease and cellular stress adaptation [9]. The presence of these complex interactions, including potential perturbations in the neurotransmitter release cycle pathway, could contribute to treatment resistance in RCB-III tumors. Future studies targeting these immunosuppressive mechanisms and exploring novel pathways could offer promising avenues to overcome resistance in this aggressive subtype.

These findings above emphasize the potential of MOMLIN to enable deeper understanding of complex biological mechanism correspondence to each response class, ultimately paving the way for personalized treatment strategies in cancer. MOMLIN also demonstrated the best prediction performance for unseen patients by utilizing these identified sets of network biomarkers. By identifying response-associated biomarkers, researchers can stratify patients based on their likelihood of achieving pCR or experiencing RCB to anticancer treatments, facilitating more informed treatment decisions and potentially improving patient outcomes. Moreover, the identified biomarkers could serve as valuable targets for the development of novel therapeutic interventions and new biological hypothesis generation. However, the clinical translation of multimodal biomarkers necessitates addressing the potential economic burden associated with multi-omics testing. Developing targeted biomarker panels and prioritizing key hub molecules from the large-scale candidate

multimodal network biomarkers identified by MOMLIN could be a viable strategy for reducing costs while maintaining predictive accuracy. Furthermore, ongoing advancements in sequencing and diagnostic technologies are expected to make multi-omics testing more accessible and affordable over time.

In conclusion, our study demonstrates MOMLIN's capacity to uncover nuanced molecular signatures associated with different drug-response classes in BC. By integrating multi-modal and -omics datasets, we have highlighted the complex interplay between genetic alterations, gene expression, immune infiltration and cellular pathways that contribute to treatment response and resistance. Future research in this direction holds promise for refining risk stratification, optimizing treatment selection and ultimately improving patient outcomes.

Limitations

While MOMLIN demonstrates promising results as shown, a key limitation lies in its reliance on correlation-based algorithms for multi-omics data integration. These algorithms are great at identifying associations, but they can fall short when it comes to inferring causality between different omics layers. This is a challenge faced by most current state-of-the-art methods [28, 30]. In the future iterations of MOMLIN, we aim to incorporate causal inference methodologies alongside sparse correlation algorithms to better understand the complex causal relationships within multi-omics datasets.

Key Points

- We proposed MOMLIN, a novel framework designed to integrate multimodal data and identify response-associated network biomarkers, to understand biological mechanisms and regulatory roles.
- MOMLIN employed an adaptive weighting for different data modalities and employs innovative regularization constraint to ensure robust feature selection to analyze high-dimensional omics data.
- MOMLIN demonstrates significantly improved performance compared to current state-of-the-art methods.
- MOMLIN identifies interpretable and phenotype-specific components, providing insights into the molecular mechanisms driving treatment response and resistance.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Acknowledgements

We thank Dr Yoshihiro Yamnishi and Mr Chen Yuzhou for their technical help.

Funding

This work was supported by the core research budget of Bioinformatics Institute, ASTAR.

Data availability

Supplemental information and software are available at the Bib website. Our algorithm's software is available for free download at https://github.com/mamun41/MOMLIN_software/tree/main

References

- Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol* 2017;**18**:83.
- Rashid MM, Hamano M, Iida M. et al. Network-based identification of diagnosis-specific trans-omic biomarkers via integration of multiple omics data. *Biosystems* 2024;**236**:105122. <https://doi.org/10.1016/j.biosystems.2024.105122>.
- Zhu B, Song N, Shen R. et al. Integrating clinical and multiple omics data for prognostic assessment across human cancers. *Sci Rep* 2017;**7**:16954. <https://doi.org/10.1038/s41598-017-17031-8>.
- Aly HA. Cancer therapy and vaccination. *J Immunol Methods* 2012;**382**:1–23.
- Debela DT. et al. New approaches and procedures for cancer treatment: current perspectives. *SAGE Open Med* 2021;**9**:20503121211034366.
- Rauf A, Abu-Izneid T, Khalil AA. et al. Berberine as a potential anticancer agent: a comprehensive review. *Molecules* 2021;**26**:7368. <https://doi.org/10.3390/molecules26237368>.
- Islam MR, Islam F, Nafady MH. et al. Natural small molecules in breast cancer treatment: understandings from a therapeutic viewpoint. *Molecules* 2022;**27**:2165. <https://doi.org/10.3390/molecules27072165>.
- Emran TB, Shahriar A, Mahmud AR. et al. Multidrug resistance in cancer: understanding molecular mechanisms. *Front Oncol* 2022;**12**:891652. <https://doi.org/10.3389/fonc.2022.891652>.
- Sammur SJ, Crispin-Ortuzar M, Chin SF. et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature* 2022;**601**:623–9. <https://doi.org/10.1038/s41586-021-04278-5>.
- Zhang A, Miao K, Sun H. et al. Tumor heterogeneity reshapes the tumor microenvironment to influence drug resistance. *Int J Biol Sci* 2022;**18**:3019–33. <https://doi.org/10.7150/ijbs.72534>.
- Karczewski KJ, Snyder MP. Integrative omics for health and disease. *Nat Rev Genet* 2018;**19**:299–310.
- In GK. et al. Multi-omic profiling reveals discrepant immunogenic properties and a unique tumor microenvironment among melanoma brain metastases. *NPJ Precis Oncol* 2023;**7**:120. <https://doi.org/10.1038/s41698-023-00471-z>.
- Denkert C, Untch M, Benz S. et al. Reconstructing tumor history in breast cancer: signatures of mutational processes and response to neoadjuvant chemotherapy (small star, filled). *Ann Oncol* 2021;**32**:500–11. <https://doi.org/10.1016/j.annonc.2020.12.016>.
- Lesurf R, Griffith OL, Griffith M. et al. Genomic characterization of HER2-positive breast cancer and response to neoadjuvant trastuzumab and chemotherapy-results from the ACOSOG Z1041 (alliance) trial. *Ann Oncol* 2017;**28**:1070–7. <https://doi.org/10.1093/annonc/mdx048>.
- Choi JH, Yu J, Jung M. et al. Prognostic significance of TP53 and PIK3CA mutations analyzed by next-generation sequencing in breast cancer. *Medicine (Baltimore)* 2023;**102**:e35267. <https://doi.org/10.1097/MD.00000000000035267>.
- Simeoni O, Piras V, Tomita M. et al. Tracking global gene expression responses in T cell differentiation. *Gene* 2015;**569**:259–66. <https://doi.org/10.1016/j.gene.2015.05.061>.
- Piras V, Hayashi K, Tomita M. et al. Enhancing apoptosis in TRAIL-resistant cancer cells using fundamental response rules. *Sci Rep* 2011;**1**:144. <https://doi.org/10.1038/srep00144>.
- Miseti H, Keddar MR, Jeannon JP. et al. Mechanistic insights into the interactions between cancer drivers and the tumour immune microenvironment. *Genome Med* 2023;**15**:40. <https://doi.org/10.1186/s13073-023-01197-0>.
- Son B, Lee S, Youn HS. et al. The role of tumor microenvironment in therapeutic resistance. *Oncotarget* 2017;**8**:3933–45. <https://doi.org/10.18632/oncotarget.13907>.
- Wang C, Lye X, Kaalia R. et al. Deep learning and multi-omics approach to predict drug responses in cancer. *BMC Bioinformatics* 2021;**22**:632. <https://doi.org/10.1186/s12859-022-04964-9>.
- Li F, Yin J, Lu M. et al. ConSIG: consistent discovery of molecular signature fromOMIC data. *Brief Bioinform* 2022;**23**:bbac253. <https://doi.org/10.1093/bib/bbac253>.
- Yang Q, Li B, Tang J. et al. Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. *Brief Bioinform* 2020;**21**:1058–68. <https://doi.org/10.1093/bib/bbz049>.
- Picard M, Scott-Boyer MP, Bodein A. et al. Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J* 2021;**19**:3735–46. <https://doi.org/10.1016/j.csbj.2021.06.030>.
- Dong Z, Zhang N, Li C. et al. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC Cancer* 2015;**15**:489. <https://doi.org/10.1186/s12885-015-1492-6>.
- Menden MP, Iorio F, Garnett M. et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One* 2013;**8**:e61318. <https://doi.org/10.1371/journal.pone.0061318>.
- Basu A, Bodycombe NE, Cheah JH. et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* 2013;**154**:1151–61. <https://doi.org/10.1016/j.cell.2013.08.003>.
- Adam G, Rampásek L, Safikhani Z. et al. Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precis Oncol* 2020;**4**:19. <https://doi.org/10.1038/s41698-020-0122-1>.
- Singh A, Shannon CP, Gautier B. et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 2019;**35**:3055–62. <https://doi.org/10.1093/bioinformatics/bty1054>.
- Wang T, Shao W, Huang Z. et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat Commun* 2021;**12**:3445. <https://doi.org/10.1038/s41467-021-23774-w>.
- Argelaguet R, Arnol D, Bredikhin D. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* 2020;**21**:111. <https://doi.org/10.1186/s13059-020-02015-1>.
- Rodosthenous T, Shahrezaei V, Evangelou M. Integrating multi-OMICS data through sparse canonical correlation analysis for the prediction of complex traits: a comparison study. *Bioinformatics* 2020;**36**:4616–25. <https://doi.org/10.1093/bioinformatics/btaa530>.
- Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol* 2009;**8**:Article28. <https://doi.org/10.2202/1544-6115.1470>.
- Jeong D, Koo B, Oh M. et al. GOAT: gene-level biomarker discovery from multi-omics data using graph Attention neural network for eosinophilic asthma subtype. *Bioinformatics* 2023;**39**:btad582. <https://doi.org/10.1093/bioinformatics/btad582>.
- Hu W, Lin D, Cao S. et al. Adaptive sparse multiple canonical correlation analysis with application to imaging (epi)genomics study of schizophrenia. *IEEE Trans Biomed Eng* 2018;**65**:390–9. <https://doi.org/10.1109/TBME.2017.2771483>.

35. Hanzelmann S, Castelo R, Guinney J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 2013;**14**:7.
36. Sotiriou C, Wirapati P, Loi S. et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 2006;**98**:262–72. <https://doi.org/10.1093/jnci/djj052>.
37. Desmedt C, Haibe-Kains B, Wirapati P. et al. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin Cancer Res* 2008;**14**:5158–65. <https://doi.org/10.1158/1078-0432.CCR-07-4756>.
38. Danaher P, Warren S, Dennis L. et al. Gene expression markers of tumor infiltrating leukocytes. *J Immunother Cancer* 2017;**5**:18. <https://doi.org/10.1186/s40425-017-0215-8>.
39. Charoentong P, Finotello F, Angelova M. et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep* 2017;**18**:248–62. <https://doi.org/10.1016/j.celrep.2016.12.019>.
40. Jiang P, Gu S, Pan D. et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat Med* 2018;**24**:1550–8. <https://doi.org/10.1038/s41591-018-0136-1>.
41. D'Eustachio P. Reactome knowledgebase of human biological pathways and processes. *Methods Mol Biol* 2011;**694**:49–61.
42. Schaefer CF, Anthony K, Krupa S. et al. PID: the pathway interaction database. *Nucleic Acids Res* 2009;**37**:D674–9. <https://doi.org/10.1093/nar/gkn653>.
43. Liberzon A, Subramanian A, Pinchback R. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011;**27**:1739–40. <https://doi.org/10.1093/bioinformatics/btr260>.
44. Du L. et al. Identifying diagnosis-specific genotype–phenotype associations via joint multitask sparse canonical correlation analysis and classification. *Bioinformatics* 2020;**36**:i371–9. <https://doi.org/10.1093/bioinformatics/btaa434>.
45. Hao X, Li C, du L. et al. Mining outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimer's disease. *Sci Rep* 2017;**7**:44272. <https://doi.org/10.1038/srep44272>.
46. Shi WJ, Zhuang Y, Russell PH. et al. Unsupervised discovery of phenotype-specific multi-omics networks. *Bioinformatics* 2019;**35**:4336–43. <https://doi.org/10.1093/bioinformatics/btz226>.
47. Duan R, Gao L, Gao Y. et al. Evaluation and comparison of multi-omics data integration methods for cancer subtyping. *PLoS Comput Biol* 2021;**17**:e1009224. <https://doi.org/10.1371/journal.pcbi.1009224>.
48. Ponzetti M, Capulli M, Angelucci A. et al. Non-conventional role of haemoglobin beta in breast malignancy. *Br J Cancer* 2017;**117**:994–1006. <https://doi.org/10.1038/bjc.2017.247>.
49. Yang X, Tang Z. Role of gasdermin family proteins in cancers (review). *Int J Oncol* 2023;**63**:100. <https://doi.org/10.3892/ijo.2023.5548>.
50. Chen Z, Yao N, Zhang S. et al. Identification of critical radioreistance genes in esophageal squamous cell carcinoma by whole-exome sequencing. *Ann Transl Med* 2020;**8**:998. <https://doi.org/10.21037/atm-20-5196>.
51. Zhou Y, Zhang Y, Zhao D. et al. TTD: therapeutic target database describing target druggability information. *Nucleic Acids Res* 2024;**52**:D1465–77. <https://doi.org/10.1093/nar/gkad751>.
52. Li F, Yin J, Lu M. et al. DrugMAP: molecular atlas and pharmacological information of all drugs. *Nucleic Acids Res* 2023;**51**:D1288–99. <https://doi.org/10.1093/nar/gkac813>.
53. Závěský L, Jandáková E, Weinberger V. et al. Human endogenous retroviruses (HERVs) in breast cancer: altered expression pattern implicates divergent roles in carcinogenesis. *Oncology* 2024;**102**:1–10. <https://doi.org/10.1159/000538021>.
54. van der Wiel AMA, Schuitmaker L, Cong Y. et al. Homologous recombination deficiency scar: mutations and beyond-implications for precision oncology. *Cancers (Basel)* 2022;**14**:4157. <https://doi.org/10.3390/cancers14174157>.
55. Morisaki T, Kubo M, Umebayashi M. et al. Neoantigens elicit T cell responses in breast cancer. *Sci Rep* 2021;**11**:13590. <https://doi.org/10.1038/s41598-021-91358-1>.
56. Levine KM, Ding K, Chen L. et al. FGFR4: a promising therapeutic target for breast cancer and other solid tumors. *Pharmacol Ther* 2020;**214**:107590. <https://doi.org/10.1016/j.pharmthera.2020.107590>.
57. Ali H, Provenzano E, Dawson SJ. et al. Association between CD8+ T-cell infiltration and breast cancer survival in 12 439 patients. *Ann Oncol* 2014;**25**:1536–43. <https://doi.org/10.1093/annonc/mdu191>.
58. Liu Y, Pan B, Qu W. et al. Systematic analysis of the expression and prognosis relevance of FBXO family reveals the significance of FBXO1 in human breast cancer. *Cancer Cell Int* 2021;**21**:130. <https://doi.org/10.1186/s12935-021-01833-y>.
59. Park SD, Saunders AS, Reidy MA. et al. A review of granulocyte colony-stimulating factor receptor signaling and regulation with implications for cancer. *Front Oncol* 2022;**12**:932608. <https://doi.org/10.3389/fonc.2022.932608>.
60. Aghamiri S, Zandsalimi F, Raei P. et al. Antimicrobial peptides as potential therapeutics for breast cancer. *Pharmacol Res* 2021;**171**:105777. <https://doi.org/10.1016/j.phrs.2021.105777>.
61. Chen R, Wang X, Fu J. et al. High FLT3 expression indicates favorable prognosis and correlates with clinicopathological parameters and immune infiltration in breast cancer. *Front Genet* 2022;**13**:956869. <https://doi.org/10.3389/fgene.2022.956869>.
62. Chen X, Zhang T, Su W. et al. Mutant p53 in cancer: from molecular mechanism to therapeutic modulation. *Cell Death Dis* 2022;**13**:974. <https://doi.org/10.1038/s41419-022-05408-1>.
63. Azimnasab-Sorkhabi P, Soltani-As M, Yoshinaga TT. et al. IDO blockade negatively regulates the CTLA-4 signaling in breast cancer cells. *Immunol Res* 2023;**71**:679–86. <https://doi.org/10.1007/s12026-023-09378-0>.
64. Sideris N, Dama P, Bayraktar S. et al. LncRNAs in breast cancer: a link to future approaches. *Cancer Gene Ther* 2022;**29**:1866–77. <https://doi.org/10.1038/s41417-022-00487-w>.
65. Burstein MD. et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin Cancer Res* 2015;**21**:1688–98.