OXFORD

Briefings in Bioinformatics, 2025, **26(2)**, bbaf073
https://doi.org/10.1093/bib/bbaf073
Problem Solving Protocol

BAMBI integrates biostatistical and artificial intelligence methods to improve RNA biomarker discovery

Peng Zhou[1], Zixiu Li[1], Feifan Liu[1], Euijin Kwon[1,2], Tien-Chan Hsieh[3], Shangyuan Ye[4], Shobha Vasudevan[5], Jung Ae Lee[1], Khanh-Van Tran[6], Chan Zhou[1,2,7,8,*]

[1]Department of Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, MA 01655, United States
[2]Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA 01655, United States
[3]Division of Hematology-Oncology, Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA 01655, United States
[4]Biostatistics Shared Resource, Knight Cancer Institute, Oregon Health and Science University, 2720 S Moody Ave, Portland, OR 97201, United States
[5]Brown RNA Center, Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, RI 02903, United States
[6]Division of Cardiology, Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA 01655, United States
[7]The RNA Therapeutics Institute, University of Massachusetts Chan Medical School, Worcester, MA 01655, United States
[8]UMass Cancer Center, University of Massachusetts Chan Medical School, Worcester, MA 01655, United States
*Corresponding author. University of Massachusetts Chan Medical School, Albert Sherman Center, Room AS9-1079, 368 Plantation Street, Worcester, MA 01605, United States. E-mail: chan.zhou@umassmed.edu

Abstract

RNA biomarkers enable early and precise disease diagnosis, monitoring, and prognosis, facilitating personalized medicine and targeted therapeutic strategies. However, identification of RNA biomarkers is hindered by the challenge of analyzing relatively small yet high-dimensional transcriptomics datasets, typically comprising fewer than 1000 biospecimens but encompassing hundreds of thousands of RNAs, especially noncoding RNAs. This complexity leads to several limitations in existing methods, such as poor reproducibility on independent datasets, inability to directly process omics data, and difficulty in identifying noncoding RNAs as biomarkers. Additionally, these methods often yield results that lack biological interpretation and clinical utility. To overcome these challenges, we present BAMBI (Biostatistical and Artificial-intelligence Methods for Biomarker Identification), a computational tool integrating biostatistical approaches and machine-learning algorithms. By initially reducing high dimensionality through biologically informed statistical methods followed by machine learning–based feature selection, BAMBI significantly enhances the accuracy and clinical utility of identified RNA biomarkers and also includes noncoding RNA biomarkers that existing methods may overlook. BAMBI outperformed existing methods on both real and simulated datasets by identifying individual and panel biomarkers with fewer RNAs while still ensuring superior prediction accuracy. BAMBI was benchmarked on multiple transcriptomics datasets across diseases, including breast cancer, psoriasis, and leukemia. The prognostic biomarkers for acute myeloid leukemia discovered by BAMBI showed significant correlations with patient survival rates in an independent cohort, highlighting its potential for enhancing clinical outcomes. The software is available on GitHub (https://github.com/CZhouLab/BAMBI).

Keywords: biomarker; machine learning; statistics; noncoding RNA; diagnosis and prognosis; high-dimensionality reduction

Introduction

Molecular biomarkers have been widely used in disease diagnosis and prognosis. There are four major types of molecular biomarkers at the biological level: DNA, RNA, protein, and metabolic biomarkers [1]. Among them, RNA biomarkers have several advantages: (i) RNA molecules have rapid and significant expression changes, allowing for real-time monitoring of disease status and treatment response [1–3]; (ii) RNA biomarkers provide insights into active gene expression and their roles in disease progression [3]; (iii) RNA molecules show higher tissue-of-origin specificity compared to DNA biomarkers, making them especially useful in disease diagnosis and prognosis [3]; (iv) multiple types of RNAs (including mRNAs and lncRNAs) can be detected in bodily fluids, such as blood, allowing for minimally invasive or noninvasive screening; and (v) the high specificity of RNA biomarkers advances personalized medicine, optimizes clinical trial processes, and accelerates the approval of new drugs. Furthermore, advances in next-generation RNA sequencing (RNA-seq) techniques make RNA biomarker identification and quantification more reliable than traditional microarray techniques.

Despite these advantages, the identification of RNA biomarkers from transcriptomics data remains challenging.

(i) Traditional machine learning (ML) or deep learning (DL) methods [4–6] often struggle with transcriptomics datasets, where the number of features far exceeds the number of samples, leading to overfitting and poor generalization [7–9]. Overfitting occurs when models perform well on training data but fail to generalize new, unseen data. Transcriptomics datasets typically comprise only tens to a few hundred samples due to the high cost of data generation, while the number of features (genes) often exceeds tens of thousands. This high dimensionality, combined with limited sample sizes, amplifies the risk of overfitting during ML or DL model training. The risk of overfitting is even greater when identifying noncoding RNA (ncRNA) biomarkers, as the number of ncRNAs (which can reach hundreds of thousands) far

Received: November 19, 2024. **Revised:** January 9, 2025. **Accepted:** January 26, 2025

exceeds that of protein-coding mRNAs (tens of thousands). This challenge is particularly critical in rare diseases, where obtaining large, high-quality datasets is especially difficult.

(ii) The biomarkers identified by current accuracy-driven methods face challenges in clinical translation. These methods, particularly DL-based approaches, often prioritize accuracy in the trade-off between accuracy and feature number. As a result, they frequently identify large biomarker panels to maximize predictive performance [6, 10–14]. However, large biomarker panels may increase the risk of overfitting and clinical testing costs, reduce assay feasibility, and hinder the interpretability and usability of biomarkers in routine clinical settings. These factors collectively delay clinical decision-making and limit the practicality of integrating such biomarkers into routine diagnostics.

(iii) Most current tools require preprocessed expression tables as input and lack the capability to handle raw RNA-seq data directly. This limitation poses a significant barrier for clinicians and researchers without specialized bioinformatics training, reducing the accessibility and applicability of these methods in diverse research and clinical environments.

(iv) Existing methods often overlook ncRNAs as potential biomarkers despite evidence suggesting that ncRNAs, including long noncoding RNAs (lncRNAs), comprise the majority of the human genome [15] and exhibit a higher degree of disease-specific expression compared to protein-coding mRNAs [16, 17]. Therefore, existing methods restrict opportunities to develop novel ncRNA-based diagnostics and treatments.

(v) Lastly, existing methods frequently lack sufficient insights into the biological relevance of their findings, limiting understanding of underlying disease mechanisms. Without comprehensive biological interpretability, the clinical relevance of identified biomarkers remains unclear, hindering their adoption in personalized medicine and therapeutic development.

To address these multifaceted challenges, we developed BAMBI (Biostatistics and Artificial-Intelligence-integrated Method for Biomarker Identification), a computational tool designed to enhance RNA biomarker discovery for disease diagnosis and prognosis. BAMBI mitigates the high dimensionality of transcriptomics data by employing biologically informed statistical techniques to retain only the most informative genes relevant to the disease, thereby reducing the risk of overfitting and enhancing model generalizability across different datasets and disease contexts. Following dimensionality reduction, BAMBI utilizes ML-based feature selection to identify a minimal yet highly predictive gene set of biomarkers, ensuring high prediction accuracy while maintaining simplicity and clinical feasibility. This approach addresses the issue of large biomarker panels, making the identified biomarkers more practical for clinical implementation.

Furthermore, BAMBI integrates raw RNA-seq preprocessing steps into its workflow, eliminating the need for preprocessed expression tables and making the tool accessible to users without specialized bioinformatics training. This feature democratizes biomarker discovery, facilitating broader clinical and research applications. Additionally, BAMBI is specifically designed to identify both coding and noncoding RNA biomarkers, accommodating the vast number of ncRNAs and enhancing the discovery of novel disease-specific biomarkers. By providing comprehensive visualization of identified biomarkers and their functional interpretations, BAMBI ensures that the findings are not only statistically robust but also biologically meaningful and easy for clinical implementation.

Our comparative analysis demonstrates that BAMBI outperforms existing methods in multiple metrics across multiple transcriptomics datasets spanning diseases such as breast cancer, psoriasis, and leukemia. Notably, the prognostic biomarkers for acute myeloid leukemia (AML) identified by BAMBI showed significant correlations with patient survival rates in an independent cohort, underscoring its potential to enhance clinical outcomes.

In summary, BAMBI represents a significant advancement in RNA biomarker discovery by effectively addressing the high dimensionality and clinical translation challenges inherent in analyzing transcriptomics datasets. Its ability to integrate raw RNA-seq data processing, robust feature selection, and the inclusion of noncoding RNAs makes BAMBI a versatile and powerful tool for advancing personalized medicine and improving clinical diagnostics.

Materials and methods

Collection of transcriptomics datasets (RNA-seq and microarray)

We included two RNA-seq datasets and two microarray datasets for this study. The two RNA-seq datasets consist of one breast cancer dataset obtained from the The Cancer Genome Atlas Program (TCGA) database [18] and one psoriasis dataset from the National Center for Biotechnology Information (NCBI) Gene Expression Omics (GEO) database (accession number: GSE54456 [19]). The TCGA breast cancer dataset includes transcriptomics data in Binary Alignment Map (BAM) format of 116 solid ductal and lobular neoplasms biospecimens and 112 adjacent normal solid tissue samples from the TCGA-Breast Invasive Carcinoma (TCGA-BRCA) project [18]. These data in BAM format were converted into FASTQ format using the bamtofastq tool in the BIOBAMBAM2 software [20], preparing them for downstream analyses. The psoriasis dataset includes the RNA-seq data of skin biospecimens from 92 psoriatic patients and 82 healthy individuals.

Two microarray datasets were obtained from previous publications [21, 22]. One comprises the transcriptomics data of colon tissues from 40 patients and 22 healthy controls [21], and the other comprises the transcriptomics data of prostate tissues from 52 patients and 50 healthy controls [22]. We preprocessed the microarray data tables to align with the input format requirements of BAMBI.

Gene annotation

We used LncBook Version 2.0 [23] as the reference gene annotations when mapping the RNA-seq data to the human reference genome (version: hg38). It contains the genomic location annotations for 19 957 coding genes and 101 293 lncRNA genes.

Preprocessing RNA-seq data in BAMBI

RNA-seq reads were mapped to the reference genome using the HISAT2 tool [24]. Subsequently, the HTSeq tool [25] was used to quantify the mapped alignments. Finally, quantified read counts mapped to each RNA gene were used to calculate the expression levels of RNA genes in Fragments Per Kilobase of transcript per Million mapped reads (FPKM). Prior to applying each tool, we converted the gene expression profiles into the required input format for each tool.

Feature selection embedded in BAMBI

BAMBI combines biologically informed statistical methods with ML methods to select features. This integrated feature selection approach effectively reduces data dimensionality, simplifies the dataset, and minimizes the data volume required for effective

model training. Details are provided in the Supplementary Materials and Methods.

Heatmap generation

We used the R package "pheatmap" with the default settings to generate the heatmap for this study.

Co-expression network analysis

We constructed the co-expression networks for lncRNAs using the "mcxarray" program in the Markov Clustering (MCL)–edge network analysis tool [26] (http://micans.org/mcl/) based on Spearman correlation. In this study, we selected a Spearman correlation cutoff of 0.8 to balance the number of singletons and the median node degree as recommended by the MCL tool [26] and following similar principles used in previous studies [27]. This threshold was empirically validated to maintain significant biological relationships while preventing over-clustering, ensuring that the resulting network clusters are both meaningful and functionally interpretable.

Gene Ontology analysis

We used the Database for Annotation, Visualization, and Integrated Discovery (DAVID) Functional Annotation platform (http://david.abcc.ncifcrf.gov/) [28, 29] to perform the Gene Ontology (GO) enrichment analysis for the protein-coding genes identified in each co-expression cluster. Only protein-coding genes with FPKM > 1 in at least one sample from the entire dataset were used as the background for GO enrichment analyses.

Survival analysis

Survival analysis was conducted by using the Python package "lifelines." The *P*-value was calculated using the log-rank test.

Performance comparison with existing methods

We compared BAMBI's performance with three existing computational tools—BioDiscML [4], ILRC [5], and ECMarker [6]—using RNA-seq, microarray, and simulated datasets. For RNA-seq and microarray data, we evaluated the detection of both single and panel biomarkers through cross-validation strategies across multiple datasets. A "single biomarker" refers to a biomarker composed of a single gene, as opposed to a panel of biomarkers that contain multiple genes. Simulated datasets were specifically designed to assess robustness and performance under varying sample sizes, including scenarios with small sample sizes. The detailed procedures for method comparisons, simulated dataset generation, and performance evaluations are described in Supplementary Methods and Materials.

Results

Overview of BAMBI's computational workflow

The clinical utility of RNA biomarkers in disease diagnosis and prognosis includes two types of RNA biomarkers: single biomarker (composed of a single gene) and panel of biomarkers (consisting of multiple genes). Typically, RNA biomarker detection involves preprocessing transcriptomic data to obtain expression profiles, followed by the identification of candidate RNA biomarkers (Fig. 1a).

Here, we introduce BAMBI (Fig. 1b), a streamlined pipeline and novel method that integrates raw RNA-seq data preprocessing with RNA biomarker identification. This pipeline is adept at identifying both individual biomarkers and panels of minimal, highly predictive biomarkers from RNA-seq or microarray data, encompassing both coding and noncoding RNAs.

BAMBI employs a four-phase process enhanced by several innovative approaches (please see Materials and Methods for details) to address key challenges in biomarker discovery as outlined in the Introduction.

- Phase 1: Data preprocessing—BAMBI preprocesses input transcriptomic data (RNA-seq or microarray), including gene expression quantification and normalization, to generate normalized gene expression profiles for downstream analysis.
- Phase 2: Biologically informed statistical-based feature selection—Unlike conventional ML feature selection, which is typically driven by accuracy, BAMBI is specifically tailored to meet the requirements of detecting biologically informative biomarkers, focusing on both predictive accuracy and biological relevance. Thus, BAMBI first utilizes a suite of biologically informed statistical methods to reduce data dimensionality by excluding genes that are not biologically meaningful. These statistical analyses include differential expression analysis, fold-change analysis, and filtering lowly expressed genes and genes with significant expression distribution overlaps.
  - *Differential expression analysis*: Differentially expressed (DE) genes are more likely to play important functional roles, so this step ensures that retained DE gene features are functionally associated with the disease under study.
  - *Fold-change* analysis: Fold change, defined as the ratio of expression levels between conditions, provides a straightforward measure of gene expression changes, helping to identify significant alterations. Genes with large fold changes are often more likely to be functionally significant and represent robust signals rather than noise. Fold-change analysis helps prioritize genes with substantial expression differences, enhancing the reliability of the biomarkers for practical use.
  - *Filtering extremely low expression genes*: Genes with very low expression levels are susceptible to be transcriptional noise. By filtering out these low-abundance genes, BAMBI enhances model stability and ensures that the selected biomarkers represent reliable expression alterations rather than noise-prone outliers.
  - *Filtering genes with significant overlapping expression distributions*: Genes with significant overlap in their expression distributions between groups (e.g. healthy versus diseased) are not effective for distinguishing these groups. Removing such genes ensures that the selected features effectively distinguish between conditions, which is crucial for developing a robust biomarker panel. Unlike traditional methods, this approach provides a more comprehensive view of gene expression differences (Supplementary Fig. S1).

  This biologically driven approach ensures that retained features are biologically meaningful, improving biomarker detection and enhancing predictive power and robustness against noise.
- Phase 3: ML-based feature selection—While the previous phase focuses on selecting genes that are biologically meaningful and relevant to the disease context, this ML-based feature selection phase will further refine the gene set, emphasizing predictive genes while minimizing the risk of overfitting. It uses recursive feature elimination (RFE) along with the SHapley Additive exPlanations (SHAP) value

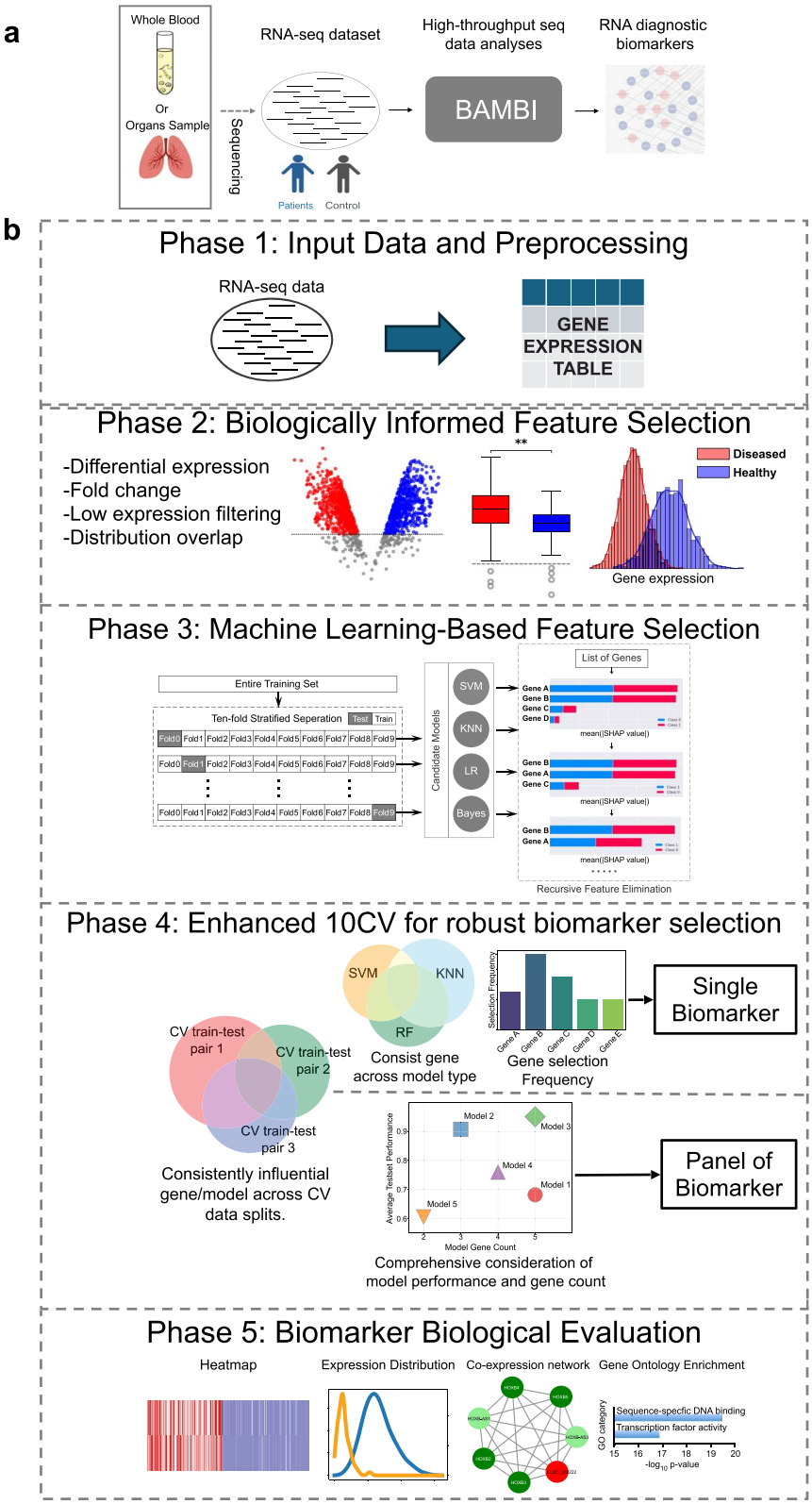


Figure 1. **Overview of the BAMBI workflow for RNA biomarker discovery.** (**a**) Illustration of BAMBI applied to RNA biomarker detection. BAMBI identifies RNA diagnostic biomarkers by analyzing transcriptomics data (RNA-seq or microarray) obtained from whole blood or tissue samples. (**b**) The BAMBI workflow comprises five phases: Phase 1 involves preprocessing input transcriptomics data, including RNA quantification and normalization, to generate normalized gene expression profiles for downstream analysis. Phase 2 employs biologically informed statistical feature selection, including differential expression analysis, fold-change analysis, and filtering of genes with low expression and genes with significant overlaps in expression distributions, to reduce data dimensionality and retain biologically relevant features. Phase 3 applies ML-based feature selection utilizing RFE with SHAP values within a 10-fold cross-validation framework to refine the gene set and emphasize predictive features. Phase 4 implements enhanced cross-validation to identify robust biomarkers, focusing on genes that are consistently influential across data splits and model types and thereby ensuring predictive reliability. BAMBI enables the provision of both single biomarkers (composed of a single gene) and panel biomarkers (consisting of multiple biomarkers). Phase 5 evaluates the biological relevance of identified biomarkers by visualizing their expression profiles with heatmaps and distribution plots as well as their co-expression clusters and enriched GO terms associated with these clusters. This provides insights into the potential functional roles of the candidate biomarkers.

[30] within a 10-fold cross-validation (10-CV) framework. Together, these two phases effectively reduce data dimensionality, simplify the dataset, and decrease the amount of data required for effective model training.

- Phase 4: Robust biomarker selection—Following feature selection, BAMBI employs an enhanced cross-validation strategy for robust biomarker selection. It fully utilizes insights generated through 10-fold cross-validation, moving beyond traditional methods that rely solely on aggregated metrics. BAMBI leverages the comprehensive information available from 10-CV to identify both single and panel biomarkers that are consistently influential across multiple data splits and model types, focusing on biomarkers with robust predictive reliability. This strategy ensures that the identified biomarkers are predictively stable and suitable for biological validation and clinical use.
- Phase 5: Evaluating biological relevance of identified biomarkers—BAMBI evaluates the biological relevance of identified biomarkers by visualizing their expression patterns (e.g. heatmaps, expression distribution plots, box plots), co-expression networks, and enriched GO terms. The co-expression network cluster identified biomarkers with genes sharing common biological functions, while GO enrichment provides insights into the potential functional roles of these biomarkers in the studied disease. This phase further assesses whether the identified biomarkers are functionally relevant to the diseases under study.

This multiphase approach balances biological relevance, predictive accuracy, and robustness, enabling BAMBI to effectively address the challenges of high-dimensional transcriptomic datasets with limited sample sizes. This algorithm's detailed pseudocode and implementation are provided in the Supplementary Materials to ensure reproducibility.

BAMBI outperforms existing methods on both real and simulated datasets

We compared the performance of BAMBI with three state-of-the-art computational tools: BioDiscML [4], ILRC [5], and ECMarker [6], using both real and simulated datasets.

First, we compared the relative performance of these four methods using both real RNA-seq and microarray datasets (Fig. 2a) in detecting two different types of biomarkers: single biomarker and panel of multiple gene biomarker (please see Materials and Methods for details). For each RNA-seq dataset, the relative performances of both lncRNA and mRNA were evaluated separately, resulting in six scenarios for performance comparison (Fig. 2a). We assessed the performance regarding specificity versus sensitivity and precision versus recall. Across all scenarios, BAMBI demonstrated superior relative performance in discovering single biomarkers and panel biomarkers (Fig. 2b and c, Supplementary Figs S2–S5, and Supplementary Tables S1 and S2). Additionally, BAMBI had the highest balanced accuracy for both single biomarkers (Fig. 2d) and panel biomarkers (Fig. 2e). While BAMBI achieves higher relative performance than other tools, its identified panel biomarkers still consist of the least number of genes compared to other tools (Fig. 2e). Fewer biomarkers make them more useful in clinical testing because they simplify the clinical test, reduce costs, and target the most relevant indicators for disease, thus enhancing their practicality in a variety of healthcare settings.

Second, we evaluated the performance of BAMBI against three other methods using simulated datasets. Specifically, we generated 100 simulated datasets based on lncRNA expression profiles from the TCGA breast cancer dataset, with each dataset generated through an independent simulation with a distinct random seed. Each simulated dataset consisted of shuffled expression profiles for three groups of lncRNA genes across patient and control cohorts: targeted biomarkers, shuffled biomarkers, and nonbiomarker genes (Fig. 2f). The five putative lncRNA biomarkers identified in the breast cancer dataset (Fig. 2a) were used as predefined biomarkers in each simulated dataset. Among them, two (*HSALNG0116686* and *HSALNG0022084*) were designated as targeted biomarkers, and the other three (*HSALNG0119995*, *HSALNG0112904*, and *HSALNG0075746*) served as shuffled biomarkers. The targeted biomarkers, whose expressions were shuffled only within their original cohorts, still retained differential expressions between patient and control cohorts, ensuring they remained valid biomarkers in each simulated dataset. In contrast, the shuffled biomarkers were designed to evaluate the effectiveness of shuffling in eliminating their expression heterogeneity, and thus, we shuffled their expression profiles across cohorts to eliminate differential expression between patients and controls. Consequently, shuffled biomarkers are no longer identifiable as valid biomarkers. Additionally, 100 nonbiomarker lncRNA genes were randomly selected from the remaining lncRNA genes, and their expression profiles were shuffled across cohorts to serve as negative controls.

To evaluate BAMBI's robustness and ability to handle small sample sizes, we designed simulated datasets with sample sizes (n) ranging from 10 to 200. Subsamples of 200, 150, 100, 50, 30, 20, and 10 were randomly selected, with an equal distribution of patient and control samples in each dataset. Biomarker detection accuracy was compared across these sample sizes (Fig. 2g, Supplementary Table S3). BAMBI consistently outperformed BioDiscML, ILRC, and ECMarker across all sample sizes. While accuracy generally decreased for all methods as sample size (n) decreased, BAMBI demonstrated greater robustness and maintained a significantly higher accuracy across all sample sizes. In contrast, ILRC performed reasonably well with larger sample sizes ($n = 200$ and $n = 150$) but showed a sharp decline in accuracy as sample size decreased. BioDiscML showed relatively stable performance but consistently lagged behind BAMBI across most sample sizes. ECMarker struggled across all sample sizes, with accuracy consistently below 10%, reflecting the limitations of its DL-based approach for small-sample scenarios. These results underscore BAMBI's robustness and adaptability for biomarker detection, especially in scenarios with limited data availability, including pilot studies.

Beyond prediction accuracy, we compared these methods at both the algorithmic and application levels (Tables 1 and 2). BAMBI's unique combination of biologically informed statistical feature selection, ML-based feature selection, and enhanced cross-validation sets it apart from existing tools (Table 1). It also offers practical advantages, such as the ability to process raw RNA-seq data, a user-friendly design, and tailored for clinical applications (Table 2).

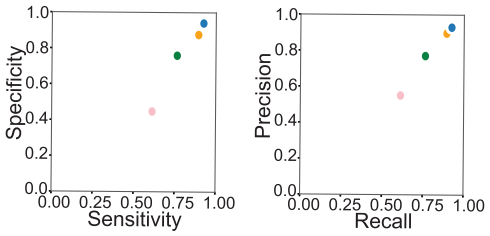
BAMBI identifies diagnostic biomarkers with high predictive power and biological significance in breast cancer

We tested BAMBI in discovering mRNA and lncRNA biomarkers for diagnosing breast cancer. BAMBI was applied to analyze the RNA-seq data of 116 biospecimens from primary ductal and lobular neoplasms as well as the RNA-seq data of 112 biospecimens from the adjacent normal solid tissues. Using BAMBI, we identified

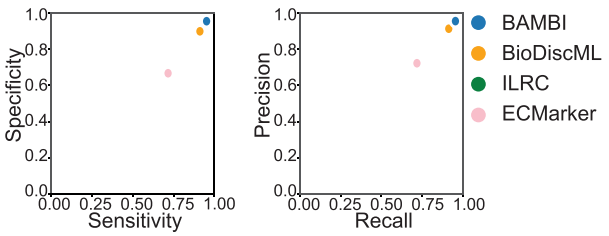
a

Dataset source	Disease	Data type	Sample size (patients)	Sample size (control)	RNA Type	Scenarios	Total number of features
TCGA	Breast cancer	RNA-seq data	116	112	mRNA	Scenario 1	19,957
					lncRNA	Scenario 2	101,293
NCBI GEO (GSE54456)	Psoriasis	RNA-seq data	92	82	mRNA	Scenario 3	19,957
					lncRNA	Scenario 4	101,293
(Alon et al., 1999)	Colon cancer	Microarray	40	22	mRNA	Scenario 5	2000
(Singh et al., 2002)	Prostate cancer	Microarray	52	50	mRNA	Scenario 6	2135

b



c



d

Singular Biomarker

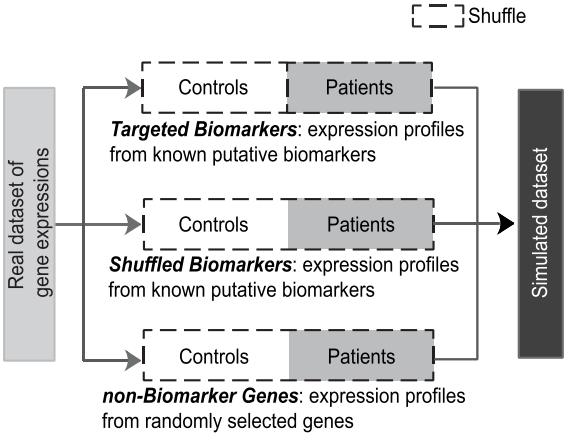
Method	Average Rank	Average Balance Accuracy
BAMBI	1.17	92.45%
BioDiscML	2.00	89.40%
ILRC	2.83	76.38%
ECMarker	4.00	61.26%

e

A Panel of Biomarkers

Method	Average Rank	Average Balance Accuracy	Average Biomarker Number
BAMBI	1.00	95.20%	2
BioDiscML	2.33	91.12%	2
ILRC	2.83	90.40%	7
ECMarker	3.75	71.68%	>1000

f



g

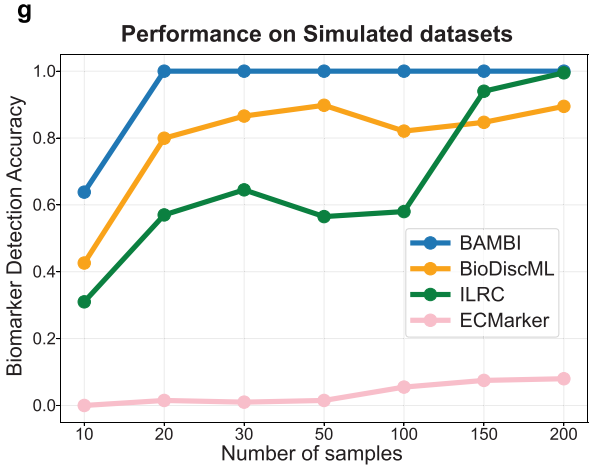


Figure 2. **BAMBI outperforms other methods on RNA-seq, microarray, and simulation data.** (**a**) Comparison of BAMBI and other methods in two RNA-seq datasets and two microarray datasets. Each RNA-seq dataset was used to identify mRNA and lncRNA biomarkers, resulting in six performance scenarios. (**b**, **c**) Scatter plots depicting the average performance differences among four methods across these six scenarios. Methods closer to the upper-right corner demonstrate superior performance. (**d**, **e**) Average performance rankings of the four methods when identifying single and panel biomarkers. (**f**) Flowchart for illustrating the generation of simulated datasets. These datasets comprised three gene categories: targeted biomarkers, shuffled biomarkers, and nonbiomarker genes. Expression profiles of targeted biomarkers were shuffled within cohorts to retain differential expression, while those of shuffled biomarkers and nonbiomarker genes were shuffled across cohorts to eliminate heterogeneity. (**g**) Performance comparison of four methods on simulated datasets with varying sample sizes. BAMBI consistently achieved higher biomarker detection accuracy across all sample sizes, demonstrating its robustness and adaptability, particularly under conditions of limited data availability. BioDiscML showed stable but lower performance; ILRC experienced a sharp decline in accuracy for datasets with smaller sample sizes; and ECMarker struggled across all scenarios, underscoring BAMBI's advantage in datasets with small sample sizes.

Table 1. Comparison of biomarker detection methods at the algorithm level.

Tools	Overall strategy	Feature selection strategies	Biomarker selection priority	Biologically informed gene filter	Optimized for ncRNAs
BAMBI	Statistical filtering + ML-based selection + enhanced cross-validation	• Differential expression • Fold-change filtering • Filtering lowly expressed genes • Filtering genes with significant expression overlapping among groups • ML recursive elimination with SHAP	• Biological relevance • Prediction accuracy • Robustness	Yes	Yes
BioDiscML	Information on gain ranking + ML-based exhaustive search	• Information gain ranking • General ML-based stepwise feature selection	• Prediction accuracy	No	No
ILRC	Clustering + L1 regularization	• Cluster-based redundancy removal • Randomized L1 regularization	• Robustness • Minimally redundant	No	No
ECMarker	Semirestricted Boltzmann machines DL model + gradient scoring	• L1 regularization	• Prediction accuracy	No	No

BAMBI combines biologically informed statistical filtering, ML-based feature selection, and enhanced cross-validation. This multifaceted approach balances biological relevance, predictive accuracy, and robustness, distinguishing it from other methods that primarily focus on prediction accuracy or general ML- or DL-based feature selection strategies. ncRNAs, noncoding RNAs.

Table 2. Comparison of biomarker detection methods at the application level.

Tools	Directly processing raw RNA-seq data?	Providing biological interpretation?	Ease of use[a]	Clinical applications[b]
BAMBI	Yes	Yes	High	High
BioDiscML	No	No	Medium	Low
ILRC	No	No	Low	Medium
ECMarker	No	No	Low	Low

BAMBI uniquely supports raw RNA-seq data input, provides biologically relevant biomarker output, and does not require extensive bioinformatics preprocessing or coding expertise. [a]Ease of use: BAMBI is user-friendly and accessible to researchers without bioinformatics expertise, whereas other tools may require advanced bioinformatics knowledge for data preprocessing and execution. Additionally, ILRC and ECMarker tools also require basic coding skills for execution. [b]Clinical applications: BAMBI aims to identify biomarkers to be minimal, robust, and interpretable. In contrast, ILRC emphasizes classification stability but does not focus on identifying a minimal number of biomarkers with interpretability. The goal of BioDiscML and ECMarker mainly focuses on achieving predictive accuracy such that they may tend to identify a large panel of biomarkers, which makes clinical implementation challenging due to complexity and interpretability issues.

both single and panel biomarkers from this dataset. Here, we focused on presenting the results of single biomarkers identified by BAMBI.

Two mRNA molecules (*TSLP* and *SPRY2*) and five lncRNA molecules (*HSALNG0022084*, *HSALNG0119995*, *HSALNG0112904*, *HSALNG0075746*, and *HSALNG0116686*) were discovered as the putative single diagnostic biomarkers for breast cancer. Each of these seven RNA biomarkers could be used to facilitate diagnosing breast cancer with a high predictive power of up to 98.7% balanced accuracy (see Fig. 3a). The expression profiles of these seven RNA biomarkers across breast cancer tumors and normal tissues, as shown in the heatmap and distribution graph (Fig. 3b and c), clearly exhibit the significantly distinct expression patterns of the discovered RNA biomarkers between breast cancer tumors and normal tissues.

The seven discovered RNA biomarkers shed light on their significant biological roles in breast cancer. *TSLP* has been identified as a biomarker and has been shown to inhibit breast cancer development through the activation of $CD4^+$ T cells [31, 32] (Fig. 3d). *SPRY2* has also been shown to negatively regulate breast cancer progression. Repressing the expression of *SPRY2* will hyperactivate the Ras/Raf/ERK pathway, promoting breast cancer progression [33–35] (Fig. 3e). The five lncRNA biomarkers are co-expressed with three groups of coding genes (Fig. 3f). Among these five lncRNA biomarkers, the co-expression network analysis and GO enrichment analysis suggest that: (i) the *HSALNG002284* was co-expressed with coding genes involved in extracellular matrix structural constituent, collagen binding, and integrin binding (Fig. 3f, left). (ii) Three of these lncRNA biomarkers (*HSALNG0119995*, *HSALNG0112904*, *HSALNG0075746*) are co-expressed with coding genes involved in calcium binding or Wnt-protein binding (Fig. 3f, middle). (iii) The other lncRNA biomarker (*HSALNG0116686*) is co-expressed with coding genes involved in angiogenesis signal transduction (Fig. 3f, right). This analysis indicates that the coding genes of these three clusters (Fig. 3f), which are co-expressed with the putative lncRNA biomarkers, have been found to contribute to breast cancer progression.

BAMBI identifies diagnostic biomarkers with high predictive power and biological significance in psoriasis

To evaluate the ability of BAMBI to discover diagnostic biomarkers in noncancerous diseases, we applied BAMBI to analyze the RNA-seq data of 174 psoriasis biospecimen from 92 patients and 82 healthy controls.

Two mRNA biomarkers (*S100A9* and *S100A7*) and two lncRNA biomarkers (*HSALNG0002446* and *HSALNG0129303*) were identified as putative biomarkers. Each of these four RNA biomarkers could be used to diagnose psoriasis with a balanced accuracy over 98% (Fig. 4a). The expression profiles of these four biomarkers (Fig. 4b and c) in heatmaps and distribution plots exhibit clear differences in expression patterns between psoriasis patients

a

Putative Biomarker	Threshold (FPKM)	Balanced Accuracy
TSLP	0.59	98.70%
SPRY2	19.66	97.50%
HSALNG0022084	0.07	95.25%
HSALNG0119995	0.22	96.90%
HSALNG0112904	0.34	95.70%
HSALNG0075746	0.09	91.66%
HSALNG0116686	0.21	96.02%

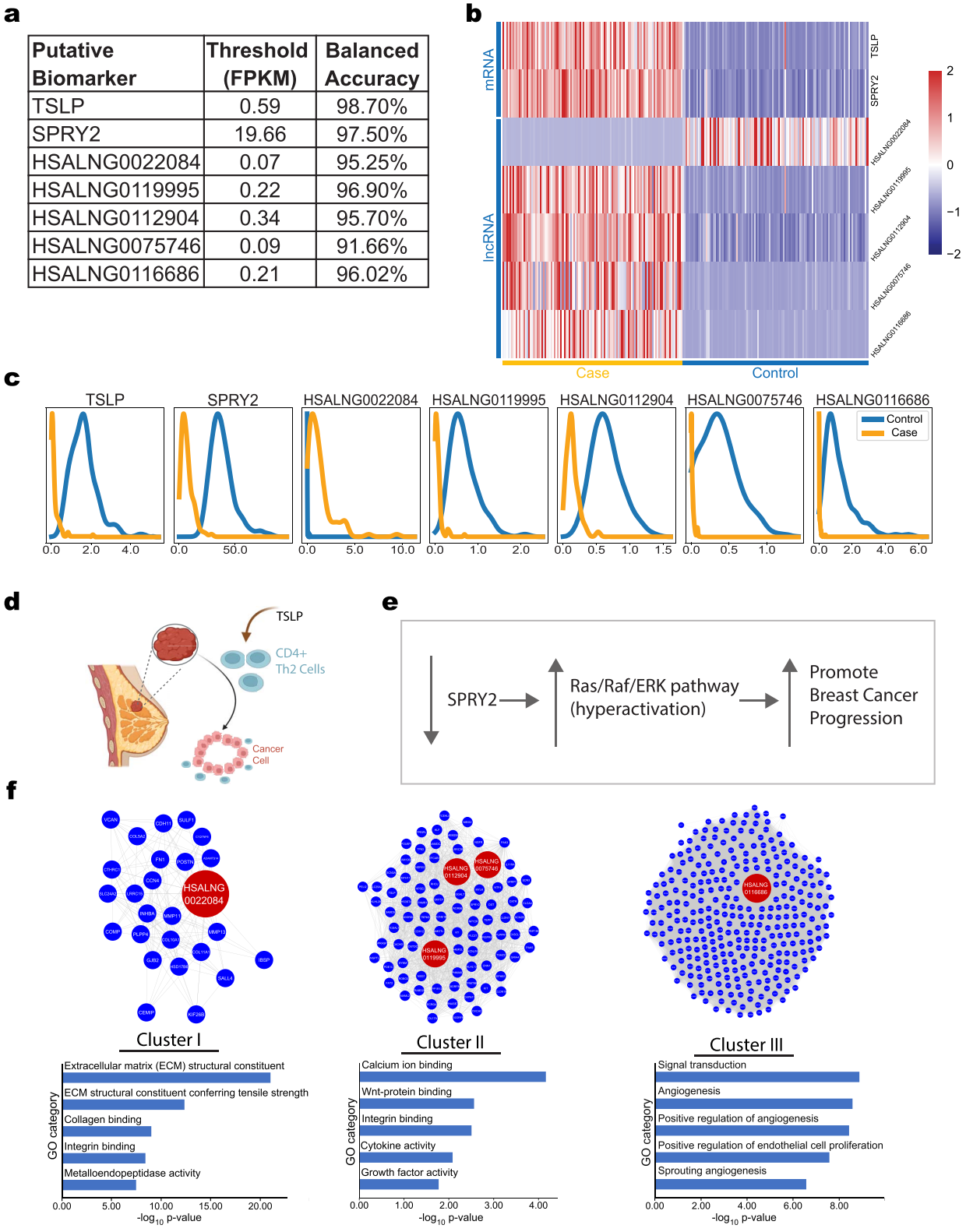


Figure 3. **BAMBI discovers the biologically informative single mRNA or lncRNA gene as the putative biomarkers for breast cancer diagnosis.** We applied BAMBI to the RNA-seq data of the tumor biospecimens and nearby normal biospecimens from a cohort of 112 breast cancer patients. BAMBI found that two protein-coding mRNAs and the five lncRNAs would be predictive for breast cancer diagnosis, each with over 91% balanced accuracy of prediction. (**a**) the prediction performance for each of the putative biomarkers in diagnosing breast cancer. (**b**) Heatmaps of the expression profiles for these seven candidate biomarkers across controls and breast cancer specimens. (**c**) Expression distributions of the seven candidate biomarkers across controls and breast cancer specimens. (**d**, **e**) Known functional roles of the putative protein-coding gene biomarkers (*TSLP* and *SPRY2*) in breast cancer. *TSLP* has been shown to inhibit breast cancer tumorigenesis through Th2 and repression of *SPRY2* expression can promote breast cancer progression through the hyperactivation of the Ras/Raf/ERK pathway. The graph **d** was created using BioRender. (**f**) The five putative lncRNA biomarkers are contained in three co-expression clusters (Clusters I, II, and III). The coding genes of Clusters I, II, and III are enriched in the GO categories significantly related to breast cancer progression. (**f**) **Left**: Cluster I contains the putative lncRNA biomarker *HSALNG0022084* and is enriched in coding genes involved in the extracellular matrix structure. (**f**) **Middle**: Cluster II contains the putative lncRNA biomarkers *HSALNG0119995*, *HSALNG0112904*, and *HSALNG0075746* and is enriched in coding genes involved in calcium ion binding and Wnt-protein binding. (**f**) **Right**: Cluster III contains the putative lncRNA biomarker *HSALNG0116686* and is enriched in coding genes involved in angiogenesis signal transduction.

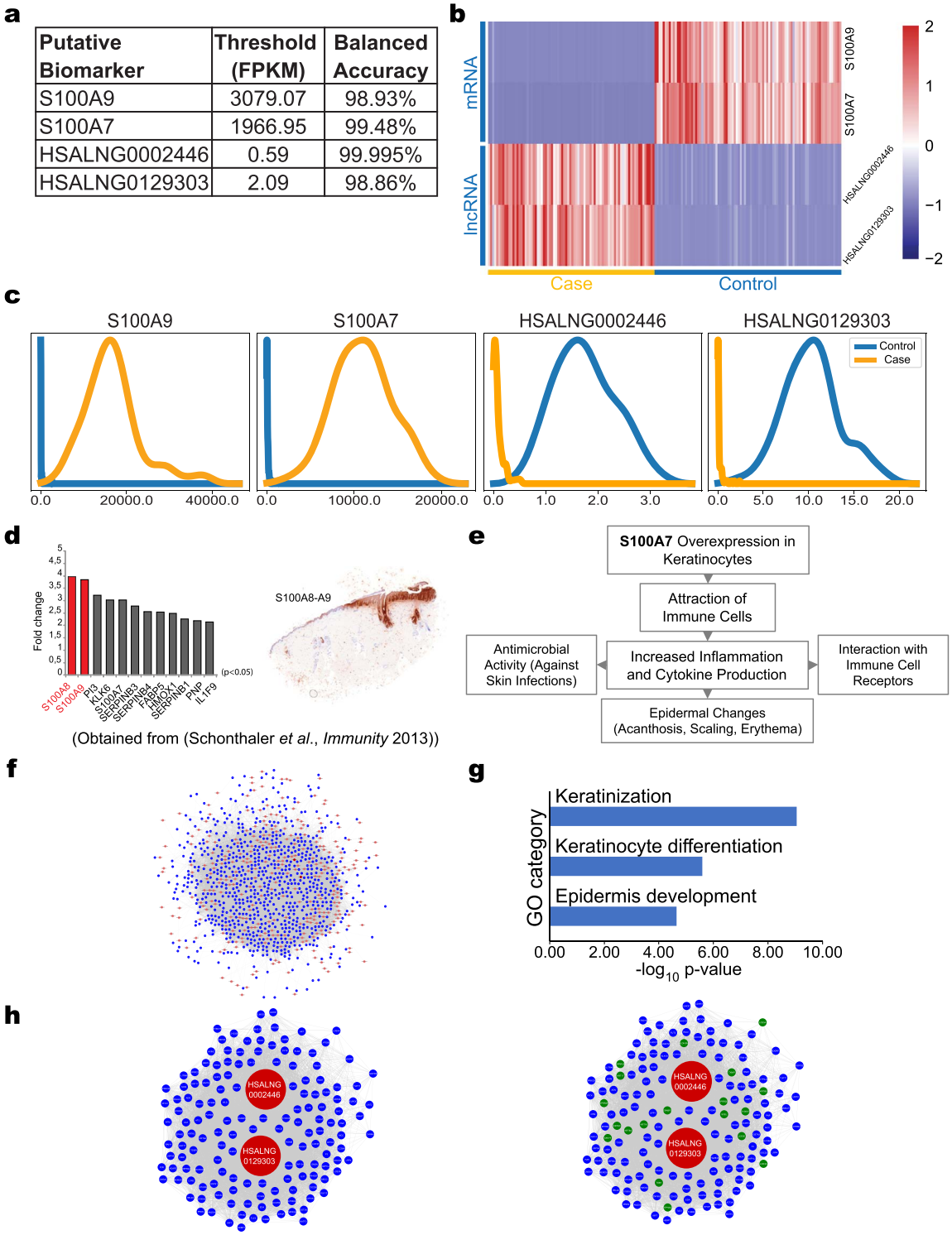


Figure 4. **BAMBI was applied to identify mRNA or lncRNA genes as putative biomarkers for diagnosing psoriasis.** We applied BAMBI to the RNA-seq data of skin biospecimens from 92 psoriasis patients and 82 healthy controls. BAMBI identified two protein-coding genes (*S100A9* and *S100A7*) and two lncRNAs (*HSALNG0002446* and *HSALNG0129303*) as predictive biomarkers for psoriasis diagnosis, each with >98% prediction power. (**a**) The prediction performance for each of the putative biomarkers in diagnosing psoriasis. (**b**) Heatmaps of the expression profiles for candidate mRNA and lncRNA biomarkers across controls and psoriasis specimens. (**c**) Expression distributions of the four candidate biomarkers across controls and psoriasis specimens. (**d**) *S100A9* is highly enriched in the psoriatic skin compared to normal skin. (**e**) *S100A7*, also known as psoriasin, is a key player in psoriasis and exhibits multifaceted roles. This protein is notably overexpressed in keratinocytes, which leads to abnormal skin cell differentiation and increased inflammation. This overexpression attracts immune cells to the skin. The accumulation of immune cells, coupled with the presence of S100A7, escalates the production of inflammatory cytokines, further amplifying inflammation and exacerbating psoriasis symptoms. (**f**) The two putative lncRNA biomarkers are contained in one co-expression cluster. (**g**) The coding genes of this co-expression cluster containing the putative lncRNA-biomarkers are enriched in the GO categories significantly related to keratinization and epidermis development. This indicates that these two putative lncRNA biomarkers may contribute to psoriasis progression through keratinization and epidermis development. (**h**) **Left**: The coding genes that are directly connected to the two putative lncRNA-biomarkers in the co-expression cluster of **f**. (**h**) **Right**: The coding genes that are involved in keratinization and epidermis development are highlighted.

and healthy controls, which are the defining characteristics of biomarkers.

The four identified RNA biomarkers can provide significant insights into the pathology of psoriasis. *S100A9* is highly enriched in psoriatic skin compared to normal skin, and this has been verified by many previous studies [36–38] (Fig. 4d). *S100A7*, also known as psoriasin, is a key player in psoriasis and exhibits multifaceted roles. This protein is notably overexpressed in keratinocytes resulting in abnormal skin cell differentiation and increased inflammation. This overexpression attracts immune cells to the skin. The accumulation of immune cells, coupled with the presence of *S100A7*, increases the production of inflammatory cytokines, further intensifying inflammation and exacerbating psoriasis symptoms [39–41] (Fig. 4e). Co-expression network analysis and GO enrichment analysis (Fig. 4f–h) suggest that the two lncRNA biomarkers are co-expressed with the coding genes involved in keratinization and epidermis development, which are key biological processes in psoriasis [42].

BAMBI identifies putative prognostic biomarkers for acute myeloid leukemia, indicative of patient survival rate

We extended the application of BAMBI to discover prognostic biomarkers for cancer. We applied the BAMBI method to analyze the gene expression profiles of patients at the onset of AML where data for the follow-up treatment outcomes were available (Fig. 5a). Then, we examined whether the putative prognostic biomarkers identified by BAMBI were associated with the survival rates of the AML patients using an independent evaluation cohort.

Because the dataset available for both the training and independent evaluation cohorts contained mRNA gene expression profiles but lacked raw RNA-seq data, we were unable to include lncRNA genes in this prognostic biomarker application. To ensure consistency in the evaluation process, we used only mRNA genes with expression profiles available in both the training and independent evaluation cohorts. BAMBI discovered nine putative biomarkers that could be identified in at least 8 out of the 10 portions in the training cohort using 10-CV strategies. Among the nine putative biomarkers, four genes (*GRAMD1B*, *DOCK1*, *CD109*, and *ALDH2*) were found to significantly indicate the survival rate of AML patients in the independent evaluation cohort ($P < .05$) (Fig. 5b–f). As determined by the Cox proportional hazards model, higher expression levels of any of these four genes were found to be associated with increased hazard rates and consequently lower survival rates in AML patients. The result from the independent cohort was consistent with the negative correlations between their expression levels with the treatment outcome in the training cohort (Supplementary Fig. S6). This suggests these four putative prognostic biomarkers serve as negative indicators of treatment outcome for AML patients. The other five biomarkers did not show significant relevance to the survival rate of AML patients from the independent cohort, which could be attributed to multiple factors. One potential reason may be that the treatment outcome (good versus poor) in the training cohort was multifactorial rather than only the patients' overall survival rate. There may also be confounding factors, such as age, relapse rate, and event-free survival. For example, younger patients generally have better overall survival than older patients, but this may not be reflected in the differences in expression of these biomarkers.

Discussion

We developed BAMBI, a comprehensive pipeline integrating statistical methods and ML to identify coding and noncoding RNA biomarkers. BAMBI addresses key challenges in RNA biomarker discovery through several innovations.

First, BAMBI effectively handles small-cohort data by combining biologically informed statistical feature selection with ML-based feature selection. This strategy selects biologically meaningful and predictive gene sets from hundreds of thousands of gene candidates, significantly reducing data dimensionality. By reducing the dimensions of high-dimensional datasets, BAMBI performs well without requiring a large training dataset. Unlike many existing tools that often underperform with limited data size and are sensitive to sample size, BAMBI demonstrates robust performance on datasets with limited sample sizes (Fig. 2g and Supplementary Figs S2–S5).

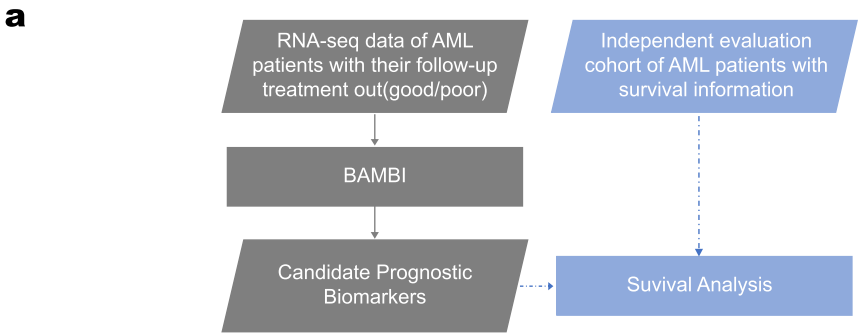
Second, BAMBI's biologically informed statistical feature selection improves the reliability and predictive power of biomarker detection. Unlike conventional ML methods that are typically accuracy-driven, BAMBI's approach aims to identify biomarkers that are both predictive and robust. This initial statistical feature selection reduces gene features, establishing a solid foundation for downstream ML-based feature selection.

Third, BAMBI employs an enhanced cross-validation strategy for robust biomarker selection. By fully utilizing insights gained from 10-fold cross-validation, BAMBI identifies genes that are consistently influential across multiple data splits and model types, emphasizing biomarkers with strong predictive reliability. This approach ensures that the identified biomarkers are consistently influential, biologically relevant, and highly interpretable, making them suitable for both biological interpretation and clinical applications.

Fourth, BAMBI demonstrates adaptability for prognostic biomarker discovery, as demonstrated with AML. BAMBI utilizes transcriptomics data from patients under on-site conditions along with their follow-up treatment outcomes, which are influenced by multiple factors such as age, survival, relapse, treatment complexity, and remission rates. Consequently, BAMBI avoids biases associated with prognostic biomarkers that only reflect survival rates, providing a more comprehensive assessment of patient prognosis.

Fifth, BAMBI is scalable to discover other types of noncoding RNAs as potential biomarkers, such as circular RNAs and microRNAs, although this study focuses on demonstrating its utility in identifying lncRNA biomarkers. It is also able to identify RNA biomarkers while combining all types of coding and noncoding RNA together.

These innovations introduced by BAMBI make it uniquely suited for practical use in both translational research and clinical settings. The following illustrate BAMBI's potential applications, ranging from early-stage pilot studies to integration into clinical workflows. First, BAMBI is well suited for pilot biomarker discovery in early-stage preclinical research, particularly when data are limited and only small cohorts are available. It serves as a valuable tool for researchers conducting pilot studies, formulating hypotheses, and laying the groundwork for larger clinical trials with more substantial data. Second, BAMBI's architecture is adaptable for handling larger sample sizes and integrating multi-omics data. As available data increase, BAMBI's dimensionality reduction techniques minimize data requirements, enabling the identification of more complex relationships with greater precision. Additionally, BAMBI's ability to handle high-dimensional data can be extended to other omics datasets, such as proteomics or DNA methylomics. Third, BAMBI is designed for ease of use and flexibility, making it accessible to a wide range of researchers, including those with limited bioinformatics expertise. Unlike other tools, BAMBI streamlines



b

mRNA gene ID	Name	Covered partitions of the 10-CV strategy in the training cohort	p-value of the survival analysis in the independent evaluation cohort
ENSG00000023171	GRAMD1B	100%	0.026
ENSG00000150760	DOCK1	90%	0.007
ENSG00000156535	CD109	90%	0.006
ENSG00000111275	ALDH2	80%	0.032

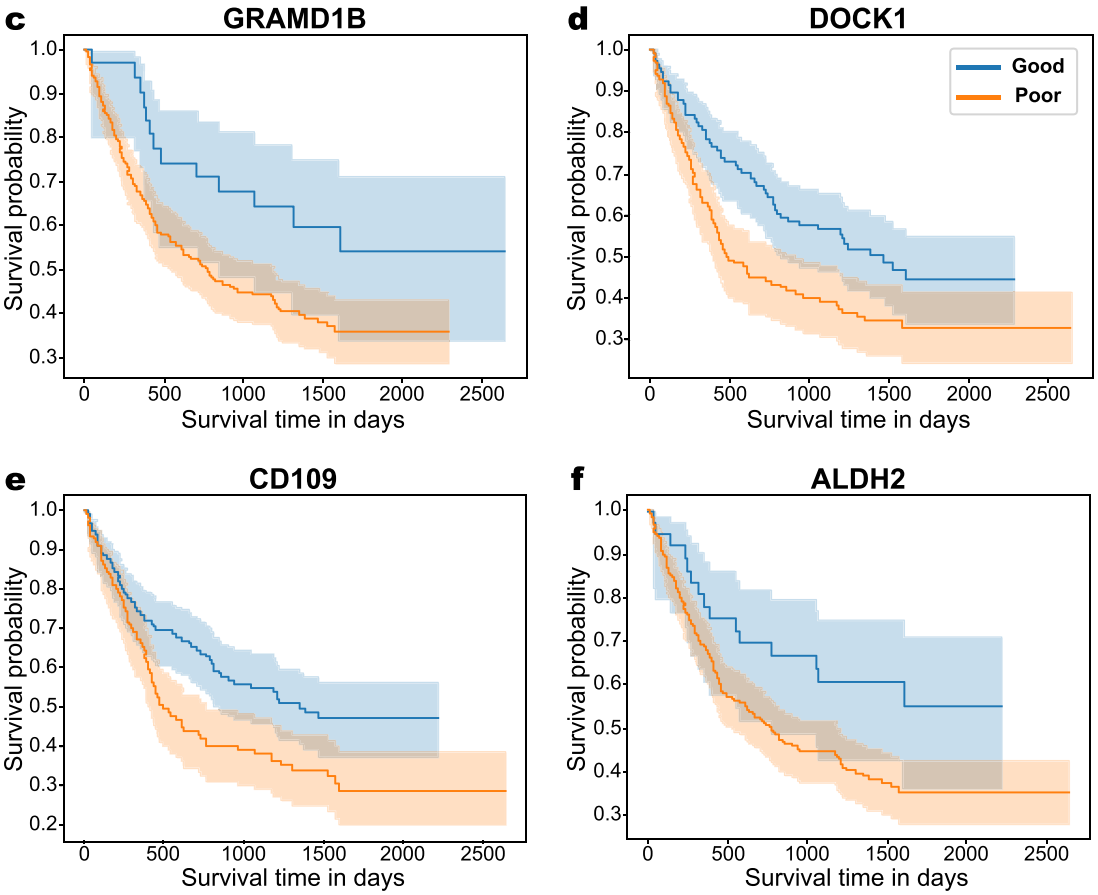


Figure 5. **BAMBI identifies putative RNA prognostic biomarkers for AML.** (**a**) Flowchart of applying BAMBI to identify RNA prognostic biomarkers for AML and evaluate them in an independent cohort with survival information. (**b**) Detailed information about each putative prognostic biomarker for AML. (**c**–**f**) Survival curves of AML patients with single putative prognostic biomarkers (c: *GRAMD1B*; d: *DOCK1*; e: *CD109*; f: *ALDH2*) in the evaluation dataset. The expression of each of these four putative prognostic biomarkers is significantly indicative of the overall survival of AML patients from the independent evaluation cohort ($P < .05$). The *P*-values were calculated using the logrank test.

the workflow, enabling broader adoption, especially in early-stage pilot studies or clinical environments where specialized bioinformatics expertise may be limited. Finally, BAMBI shows significant potential for clinical translation. For instance, the RNA biomarkers identified by BAMBI could be incorporated into diagnostic assays to guide treatment decisions and monitor therapeutic efficacy. Its capability to identify biomarkers with high predictive accuracy, combined with its adaptability to noncoding RNAs, positions BAMBI as a critical tool for advancing personalized medicine.

Future directions: BAMBI could be further enhanced to improve its application effectiveness. First, although BAMBI performs robustly with relatively small datasets, its accuracy may decline for datasets with extremely small sample sizes (e.g. <30). Future work will focus on adapting BAMBI to better handle such cases through advanced augmentation techniques or transfer learning. Second, imbalanced datasets may introduce biases during model training, which can affect generalizability. To address this, we recommend generating synthetic data to ensure balanced performance across diverse datasets. Third, while the quality of transcriptomics datasets affects the performance of all computational methods, and not BAMBI specifically, we emphasize that using high-quality datasets is essential for obtaining reliable results in any transcriptomics analysis.

Conclusion

This BAMBI method and software represent a significant advancement in identifying coding and noncoding RNA biomarkers for diagnosis and prognosis. We have demonstrated that BAMBI can discover RNA biomarkers to accurately diagnose diseases or predict treatment outcomes. It can identify both single biomarkers and biomarker panels composed of a minimal number of genes while maintaining high predictive accuracy, enhancing its applicability in clinical practice. The BAMBI tool is versatile and applicable to numerous diseases on a transcriptome-wide scale, making it a valuable asset for both translational research and clinical applications.

Key Points

- We developed BAMBI, a robust method that identifies coding and noncoding RNA biomarkers for disease diagnosis and prognosis, consistently outperforming existing methods in diverse disease contexts.
- BAMBI effectively handles small-cohort, high-dimensional transcriptomics data by using a two-step feature reduction strategy, ensuring reliable biomarker identification even for datasets with limited sample sizes.
- BAMBI's biologically informed statistical feature selection ensures that the biomarkers are both predictive and biologically meaningful, enhancing clinical interpretability.
- BAMBI's enhanced cross-validation strategy ensures that biomarkers are consistently influential, highly interpretable, and suitable for downstream clinical applications.
- BAMBI's ease of use, adaptability, and scalability make it ideal for biomarker discovery in both early-stage preclinical research and clinical applications while supporting personalized medicine and targeted therapies.

Acknowledgements

We would like to thank Dr. Arlene Ash and Dr. Honghuang Lin for their insightful suggestions on this project and manuscript.

Author contributions

Conceptualization: C.Z., P.Z.; Methodology: P.Z., C.Z., Z.L., F.L., S.Y., J.L.; Investigation: P.Z., Z.L.; Formal Analysis: P.Z., Z.L., C.Z., T.H., S.V., K.V.T.; Software: P.Z., E.K.; Validation: E.K., P.Z.; Data Curation: P.Z.; Visualization: P.Z., C.Z.; Writing–Original Draft: P.Z., C.Z.; Writing–Review & Editing: P.Z., C.Z., Z.L., F.L., T.H., S.Y., E.K., K.V.T., S.V.; Supervision: C.Z.; Project Administration: C.Z.; Funding Acquisition: C.Z.

Supplementary data

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

Conflict of Interest: None declared.

Funding

This work was supported by NIH R03DE032455-01 and NIH UL1TR001453 (to C.Z). This work was also supported by the University of Massachusetts Chan Medical School start-up funds to C.Z.

Data availability

All expression data processed from raw RNA-seq and microarray data are available in the BAMBI's GitHub repository: https://github.com/CZhouLab/BAMBI.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work, the authors used ChatGPT-4o to refine the English language in parts of the manuscript. The authors reviewed and edited the content as needed and take full responsibility for the final version of the publication.

References

1. FDA. BEST (Biomarkers, EndpointS, and other Tools) Resource. 2016.
2. Byron SA, van Keuren-Jensen KR, Engelthaler DM. *et al.* Translating RNA sequencing into clinical diagnostics: Opportunities and challenges. *Nat Rev Genet* 2016;**17**:257–71. https://doi.org/10.1038/nrg.2016.10
3. Xi X, Li T, Huang Y. *et al.* RNA biomarkers: Frontier of precision medicine for cancer. Noncoding. *RNA* 2017;**3**:9. https://doi.org/10.3390/ncrna3010009
4. Leclercq M, Vittrant B, Martin-Magniette ML. *et al.* Large-scale automatic feature selection for biomarker discovery in high-dimensional omics data. *Front Genet* 2019;**10**:452. https://doi.org/10.3389/fgene.2019.00452
5. Yu K, Xie W, Wang L. *et al.* ILRC: A hybrid biomarker discovery algorithm based on improved L1 regularization and clustering in microarray data. *BMC Bioinformatics* 2021;**22**:514. https://doi.org/10.1186/s12859-021-04443-7

6. Jin T, Nguyen ND, Talos F. *et al.* ECMarker: Interpretable machine learning model identifies gene expression biomarkers predicting clinical outcomes and reveals molecular mechanisms of human disease in early stages. *Bioinformatics* 2021;**37**:1115–24. https://doi.org/10.1093/bioinformatics/btaa935
7. Ng S, Masarone S, Watson D. *et al.* The benefits and pitfalls of machine learning for biomarker discovery. *Cell Tissue Res* 2023;**394**:17–31. https://doi.org/10.1007/s00441-023-03816-z
8. Diaz-Uriarte R, Gómez de Lope E, Giugno R. *et al.* Ten quick tips for biomarker discovery and validation analyses using machine learning. *PLoS Comput Biol* 2022;**18**:e1010357. https://doi.org/10.1371/journal.pcbi.1010357
9. Mohammed MA, Abdulkareem KH, Dinar AM. *et al.* Rise of deep learning clinical applications and challenges in omics data: A systematic review. *Diagnostics* 2023;**13**:664. https://doi.org/10.3390/diagnostics13040664
10. Perera-Bel J, Leha A, Beißbarth T. Bioinformatic methods and resources for biomarker discovery, validation, development, and integration. In: S. Badve and G. Kumar (eds), *Predictive Biomarkers in Oncology*. Cham: Springer, 2019, 123–35. https://doi.org/10.1007/978-3-319-95228-4_11
11. Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**:436–44. https://doi.org/10.1038/nature14539
12. Park MK, Lim JM, Jeong J. *et al.* Deep-learning algorithm and concomitant biomarker identification for NSCLC prediction using multi-omics data integration. *Biomolecules* 2022;**12**:1839. https://doi.org/10.3390/biom12121839
13. Kakati T, Bhattacharyya DK, Kalita JK. *et al.* DEGnext: Classification of differentially expressed genes from RNA-seq data using a convolutional neural network with transfer learning. *BMC Bioinformatics* 2022;**23**:17. https://doi.org/10.1186/s12859-021-04527-4
14. Wang A, Hai R, Rider PJ. *et al.* Noncoding RNAs and deep learning neural network discriminate multi-cancer types. *Cancers (Basel)* 2022;**14**:352. https://doi.org/10.3390/cancers14020352
15. Dunham I, Kundaje A, Aldred SF. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74. https://doi.org/10.1038/nature11247
16. Ratti M, Lampis A, Ghidini M. *et al.* MicroRNAs (miRNAs) and long non-coding RNAs (lncRNAs) as new tools for cancer therapy: First steps from bench to bedside. *Target Oncol* 2020;**15**:261–78. https://doi.org/10.1007/s11523-020-00717-x
17. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. *Annu Rev Biochem* 2012;**81**:145–66. https://doi.org/10.1146/annurev-biochem-051410-092902
18. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA. *et al.* The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013;**45**:1113–20. https://doi.org/10.1038/ng.2764
19. Li B, Tsoi LC, Swindell WR. *et al.* Transcriptome analysis of psoriasis in a large case-control sample: RNA-seq provides insights into disease mechanisms. *J Invest Dermatol* 2014;**134**:1828–38. https://doi.org/10.1038/jid.2014.28
20. Tischler G, Leonard S. Biobambam: Tools for read pair collation based algorithms on BAM files. *Source Code Biol Med* 2014;**9**:13. https://doi.org/10.1186/1751-0473-9-13
21. Alon U, Barkai N, Notterman DA. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA* 1999;**96**:6745–50. https://doi.org/10.1073/pnas.96.12.6745
22. Singh D, Febbo PG, Ross K. *et al.* Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 2002;**1**:203–9. https://doi.org/10.1016/S1535-6108(02)00030-2
23. Ma L, Cao J, Liu L. *et al.* Lncbook: A curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Res* 2019;**47**:D128–34. https://doi.org/10.1093/nar/gky960
24. Kim D, Paggi JM, Park C. *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019;**37**:907–15. https://doi.org/10.1038/s41587-019-0201-4
25. Anders S, Pyl PT, Huber W. HTSeq-A python framework to work with high-throughput sequencing data. *Bioinformatics* 2015;**31**:166–9. https://doi.org/10.1093/bioinformatics/btu638
26. Van Dongen S, Abreu-Goodger C. Using MCL to extract clusters from networks. In: van Helden J, Toussaint A, Thieffry D (eds.), *Bacterial Molecular Networks: Methods and Protocols*, pp. 281–95. New York: Springer, 2012.
27. Zhou C, York SR, Chen JY. *et al.* Long noncoding RNAs expressed in human hepatic stellate cells form networks with extracellular matrix proteins. *Genome Med* 2016;**8**:31. https://doi.org/10.1186/s13073-016-0285-0
28. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;**4**:44–57. https://doi.org/10.1038/nprot.2008.211
29. Jiao X, Sherman BT, Huang DW. *et al.* DAVID-WS: A stateful web service to facilitate gene/protein list analysis. *Bioinformatics* 2012;**28**:1805–6. https://doi.org/10.1093/bioinformatics/bts251
30. Lundberg SM, Allen PG, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017;**30**:4765–74.
31. Demehri S, Cunningham TJ, Manivasagam S. *et al.* Thymic stromal lymphopoietin blocks early stages of breast carcinogenesis. *J Clin Invest* 2016;**126**:1458–70. https://doi.org/10.1172/JCI83724
32. Boieri M, Malishkevich A, Guennoun R *et al.* CD4+ T helper 2 cells suppress breast cancer by inducing terminal differentiation. *Journal of Experimental Medicine* 2022;**219**:e20201963. https://doi.org/10.1084/jem.20201963
33. Faratian D, Sims AH, Mullen P *et al.* Sprouty 2 is an independent prognostic factor in breast cancer and may be useful in stratifying patients for trastuzumab therapy. *PLoS One* 2011;**6**:e23772. https://doi.org/10.1371/journal.pone.0023772
34. Kawazoe T, Taniguchi K. The Sprouty/Spred family as tumor suppressors: Coming of age. *Cancer Sci* 2019;**110**:1525–35. https://doi.org/10.1111/cas.13999
35. Hanafusa H, Torii S, Yasunaga T, *et al.* Sprouty1 and Sprouty2 provide a control mechanism for the Ras/MAPK signalling pathway. *Nat Cell Biol* 2002;**4**:850–58. https://doi.org/10.1038/ncb867
36. Broome A-M, Ryan D, Eckert RL. S100 protein subcellular localization during epidermal differentiation and psoriasis. *J Histochem Cytochem* 2003;**51**:675–85. https://doi.org/10.1177/002215540305100513
37. Schonthaler HB, Guinea-Viniegra J, Wculek SK. *et al.* S100A8-S100A9 protein complex mediates psoriasis by regulating the expression of complement factor C3. *Immunity* 2013;**39**:1171–81. https://doi.org/10.1016/j.immuni.2013.11.011
38. Silva de Melo BM, Veras FP, Zwicky P. *et al.* S100A9 drives the Chronification of Psoriasiform inflammation by inducing IL-23/type 3 immunity. *J Invest Dermatol* 2023;**143**:1678–1688.e8. https://doi.org/10.1016/j.jid.2023.02.026
39. Zhou X, Niu Z, Wang Y. *et al.* Advances in the pathogenesis of psoriasis: From keratinocyte perspective. *Cell Death Dis* 2022;**8**:13. https://doi.org/10.1038/s41420-021-00769-6
40. Luo M, Huang P, Pan Y. *et al.* Weighted gene coexpression network and experimental analyses identify lncRNA SPRR2C as a regulator of the IL-22-stimulated HaCaT cell phenotype through the miR-330/STAT1/S100A7 axis. *Cell Death Dis* 2021;**12**:86. https://doi.org/10.1038/s41419-020-03305-z

41. Ekman AK, Vegfors J, Eding CB. *et al.* Overexpression of psoriasin (S100A7) contributes to dysregulated differentiation in psoriasis. *Acta Derm Venereol* 2017;**97**:441–8. https://doi.org/10.2340/00015555-2596

42. Iizuka H, Takahashi H, Honma M. *et al.* Unique keratinization process in psoriasis: Late differentiation markers are abolished because of the premature cell death. *J Dermatol* 2004;**31**:271–6. https://doi.org/10.1111/j.1346-8138.2004.tb00672.x

Briefings in Bioinformatics, 2025, **26(2)**, bbaf073
https://doi.org/10.1093/bib/bbaf073
Problem Solving Protocol