SUPPLEMENTARY DATA

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

Authors: 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, USA 01655

[2]Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA 01655

[3]Division of Hematology-Oncology, Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA, USA 01655

[4]Biostatistics Shared Resource, Knight Cancer Institute, Oregon Health and Science University, Portland, OR 97239

[5]Brown RNA Center, Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, RI, USA 02903

[6]Division of Cardiology, Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA, USA 01655

[7]The RNA Therapeutics Institute, University of Massachusetts Chan Medical School, Worcester, MA, USA 01655

[8]UMass Cancer Center, University of Massachusetts Chan Medical School, Worcester, MA, USA 01655

*Correspondence: chan.zhou@umassmed.edu

Content

SUPPLEMENTARY MATERIALS AND METHODS

1. Feature selection embedded in BAMBI

BAMBI combines biologically informed statistical methods with machine learning methods to select features. This integrated feature selection approach effectively reduces data dimensionality, simplifies the dataset, and minimizes the amount of data required for effective model training.

1.1 Biologically Informed Statistical Feature Selection

We employed a suite of statistical methods tailored to the specific needs of biomarker identification:

(1.1) *Differential Expression (DE) analysis:* For RNA-seq data, we adopted the Wilcoxon-Mann-Whitney test as the default DE analysis method in BAMBI. This is because the Wilcoxon-Mann-Whitney test is a non-parametric method that has been shown to accurately determine differentially expressed genes in RNA-seq datasets with more than eight samples [1]. Compared to the conventional DE analysis methods such as DESeq2 [2] and EdgeR [3], the Wilcoxon-Mann-Whitney test shows robustness to outliers and consistent control of false positive rates at desired thresholds for population-level studies [1]. Additionally, we integrated the limma method [4] in the BAMBI package to perform DE analysis for microarray data. An adjusted p-value threshold of less than 0.05 was used for initially DE gene selection in BAMBI. If the resulting number of genes is fewer than the sample size, instead, the unadjusted p-value threshold of 0.05 is suggested to ensure an adequate number of features for downstream analysis.

(1.2) *Fold-change analysis:* In BAMBI, we calculate the fold-change score based on the median gene expression level ratios between different cohorts. Genes will be considered to have significant fold-change if having at least two-fold change between cohorts, including two-fold increasing or decreasing.

(1.3) *Filtering extremely low expression genes:* BAMBI excludes genes whose maximum expression levels are below a certain threshold: FPKM = 1.0 for coding genes and FPKM = 0.01 for lncRNAs, which are commonly expressed 10–100-fold lower than coding genes [5–7].

(1.4) *Filtering genes with significant overlapping expression-distributions:* BAMBI removes genes with significant overlap in their expression distributions between cohorts (e.g., healthy vs. diseased) to retain genes features reliably distinguishable between cohorts. Expression distributions are estimated using kernel density estimation (KDE) implemented in the Python “KDEpy” package, and the overlap area between distributions is calculated using the Python “numpy” package. Genes will be excluded if the overlap area between their distributions exceeds 50%.

1.2 Machine learning-based feature selection

By following the biologically informed statistical-based feature selection, BAMBI further incorporates a machine learning-based feature selection, focusing on enhancing prediction accuracy at the individual sample level. We integrated the following strategies into the BAMBI tool for feature selection when building machine learning models:

(2.1) *Recursive feature elimination for optimizing gene selection*: We employ recursive feature elimination (RFE) for feature selection to identify the optimal gene set, which is a widely recognized algorithm for this purpose. Initially, the model is trained on the entire gene set, and the significance of each gene for prediction is determined. Subsequently, genes deemed to be the least important are excluded and the model's prediction performance is concurrently recorded based on the remaining genes. This pruning process is repeated iteratively until only a single gene remains. Plotting these recorded performances against gene count allows a curve to be generated. The optimal gene set is determined by identifying the knee point of this curve. We have selected the SHAP algorithm [8] to evaluate the gene predictive significance for RFE. The SHAP algorithm is known for its game theory foundation and lucid interpretation, and it can be integrated with almost all conventional machine learning and deep learning models. We chose this instead of another cutting-edge machine learning interpretation method, LIME, which can sometimes exhibit instability. The performance of LIME is also contingent upon specific parameter settings, such as the neighborhood kernel width.

(2.2) *Enhanced ten-fold cross-validation for robust biomarker selection:* In BAMBI, an enhanced cross-validation strategy is used to select biomarkers by identifying genes that are consistently influential across different data splits, ensuring robustness and biological relevance. We built a structure similar to the ten-fold cross-validation (10-CV) method, a classic machine-learning resampling method which divides the dataset into 10 sub-sets to do the feature selection. Because the transcriptomics datasets are commonly unbalanced between different groups of biospecimens, we kept each sub-set containing approximately the same percentage of samples of each target class as the original dataset. We used nine sub-sets as the training set for gene selection and model training and the remaining sub-set as the testing set, resulting in 10 train-test set pairs. Then, we did the feature selection and model training on each training set and evaluated their performance on the relevant testing set.

Conventional approaches employ 10-CV for feature selection by assessing and comparing the aggregated 10-CV performance metrics of various feature selection hyperparameters. The hyperparameter that demonstrates the highest aggregated performance is then selected to guide feature selection on the entire training set, ultimately determining the final feature set. However, this may overlook valuable insights derived from the 10-CV process. For instance, a set of features consistently selected across different training and testing pairs could indicate higher robustness and predictive power. Similarly, features frequently chosen by diverse model types which are based on varying assumptions might exhibit greater heterogeneity.

To fully leverage the comprehensive information available from the 10-CV process, BAMBI employs a two-layered statistical analysis to enhance the identification of both single RNA biomarkers and panels of multiple RNA biomarkers. This approach ensures that genes consistently influential across data splits are prioritized, enhancing the robustness of the final biomarker set.

(2.3) Two-layer statistical downstream analysis of 10-CV: By applying M types of models to 10 paired training-testing sets of 10-CV process, we generated 10 × M trained candidate predictive models. Each model, through its independent feature selection process, potentially selects different gene combinations as predictors. BAMBI collects information from these 10 × M candidate models, including the predictor gene information, model type, and their relative test set performance.

Based on this information, BAMBI conducts a two-layered statistical analysis to predict single RNA biomarkers and a panel of multiple RNA biomarkers. When predicting the single biomarker, the BAMBI method hinges on three statistical criteria: the number of training-testing set pairs the gene be selected, the number of model types including the gene, and the frequency of gene presence across models. Of these, the 'covered training-testing set pairs number' is paramount, as it validates the gene's performance across diverse data splits, avoiding overfitting to specific sets. A gene's robustness is further supported if it is consistently selected by diverse models that operate on different underlying classification principles ('model type number'), suggesting greater group heterogeneity. 'Frequency' also plays a role, reinforcing the gene's discriminative power if it is selected across numerous models.

When predicting a panel of multiple RNA biomarkers, the 10 × M candidate models are aggregated based on their identical gene usage and model type, with an emphasis on the 'the number of covered training-testing set pairs' and the model's average test set performance metrics. A model is considered significantly reliable if it is selected in 70% or more of the training-testing pairs. This priority ranking ('covered pairs number' > 'average test set performance' > 'model gene count') is adjusted when a model's selection frequency is low (*e.g.*, 'covered pairs number' of one to two), shifting importance to 'average test set performance'. This flexible approach allows for the discernment of the most robust predictive models, accommodating both generalizability and performance.

(2.4) *Model selection and hyperparameter optimization:* In BAMBI, we have selected a range of machine learning classification models as potential candidates, including support vector machine (SVM), K-nearest neighbor (KNN), logistic regression, and naïve Bayes. These classifiers are prevalent in bioinformatics analyses and have consistently demonstrated commendable performance [9–11]. To select the best model training hyperparameters for each type of machine learning classification model, we applied the grid search method in BAMBI. For each model, we define an appropriate set of possible hyperparameter combinations. The grid search method exhaustively generates candidates from all possible hyperparameter combinations and evaluates the model based on the ten-fold cross-validation. The hyperparameters with the best cross-validation score were used to train the final model.

2. Comparison of methods on RNA-seq and microarray data

We compared the performance of BAMBI with current alternative methods, including BioDiscML [12], ILRC [13], and ECMarker [14], in two RNA-seq datasets for breast cancer and psoriasis [15,16] and two microarray datasets for colon cancer and prostate cancer [15–18]. Because the other three comparison methods cannot directly process RNA-seq data by default, we used the gene expression profiles generated in Phase 2 of BAMBI as the input files for these methods.

We compared the four methods for detecting two different types of biomarkers: a single biomarker and a panel of multiple RNA biomarkers. Unlike BAMBI, the other three methods were not designed to predict single biomarkers. Therefore, for the other methods, the most important feature of the best model was used to represent their single biomarkers.

Here, we utilized the cross-validation comparison strategy to evaluate the performance of each method. This is because the split of training and testing data strongly influences the performance score of the biomarker detection method, especially for small datasets. This cross-validation comparison strategy can ensure a robust and rigorous performance evaluation in small datasets.

For each dataset, we performed two independent iterations of five-fold stratified cross-validation splits. This resulted in 10 different train-test set pairs, with each training set comprising 80% of the samples and each testing set including 20% of the samples. We applied all methods to these 10 different training sets to predict putative biomarkers. Then, we evaluated the performance of these identified biomarkers using the corresponding testing set. The average performance of the 10 train-test set pairs was used to compare the performances of BAMBI and other methods.

3. Generation of simulated datasets

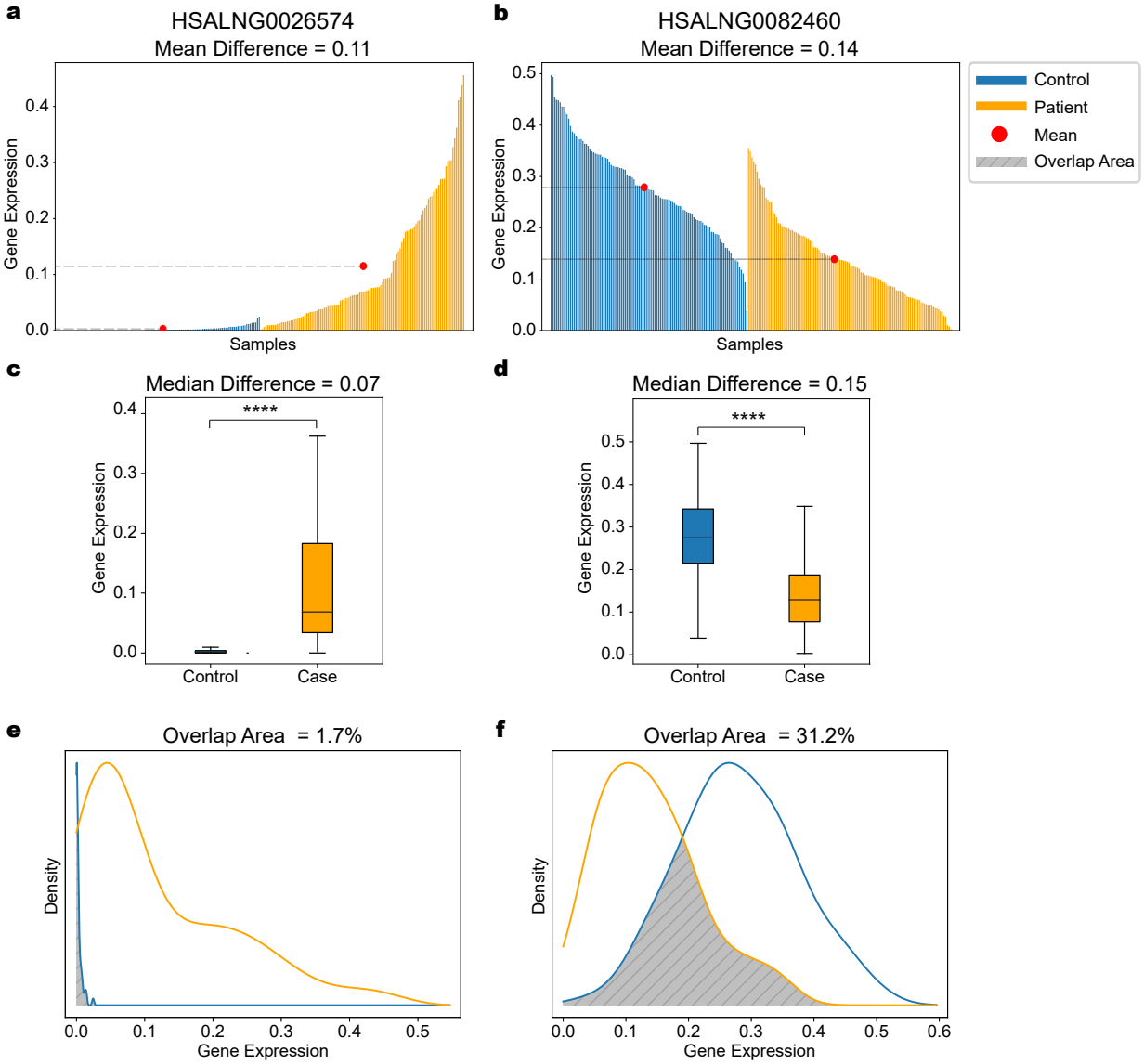
The simulated datasets were generated based on the actual gene expression profiles of patients and controls. Each simulated dataset includes gene expression profiles for three categories: targeted biomarkers, shuffled biomarkers, and non-biomarker genes. Targeted and shuffled biomarkers were selected from potential biomarkers that were significantly differentially expressed between patient and control cohorts, based on TCGA breast cancer dataset. Non-biomarker genes were randomly selected from the remaining genes. The expression profiles of targeted biomarkers were shuffled separately within patient and control cohorts, allowing them to be identified by biomarker detection methods. For shuffled biomarker and non-biomarker genes, their expression levels were shuffled across patient and control cohorts to eliminate potential heterogeneity in expression between cohorts.

4. Evaluation of biomarker detection performance on the simulated datasets

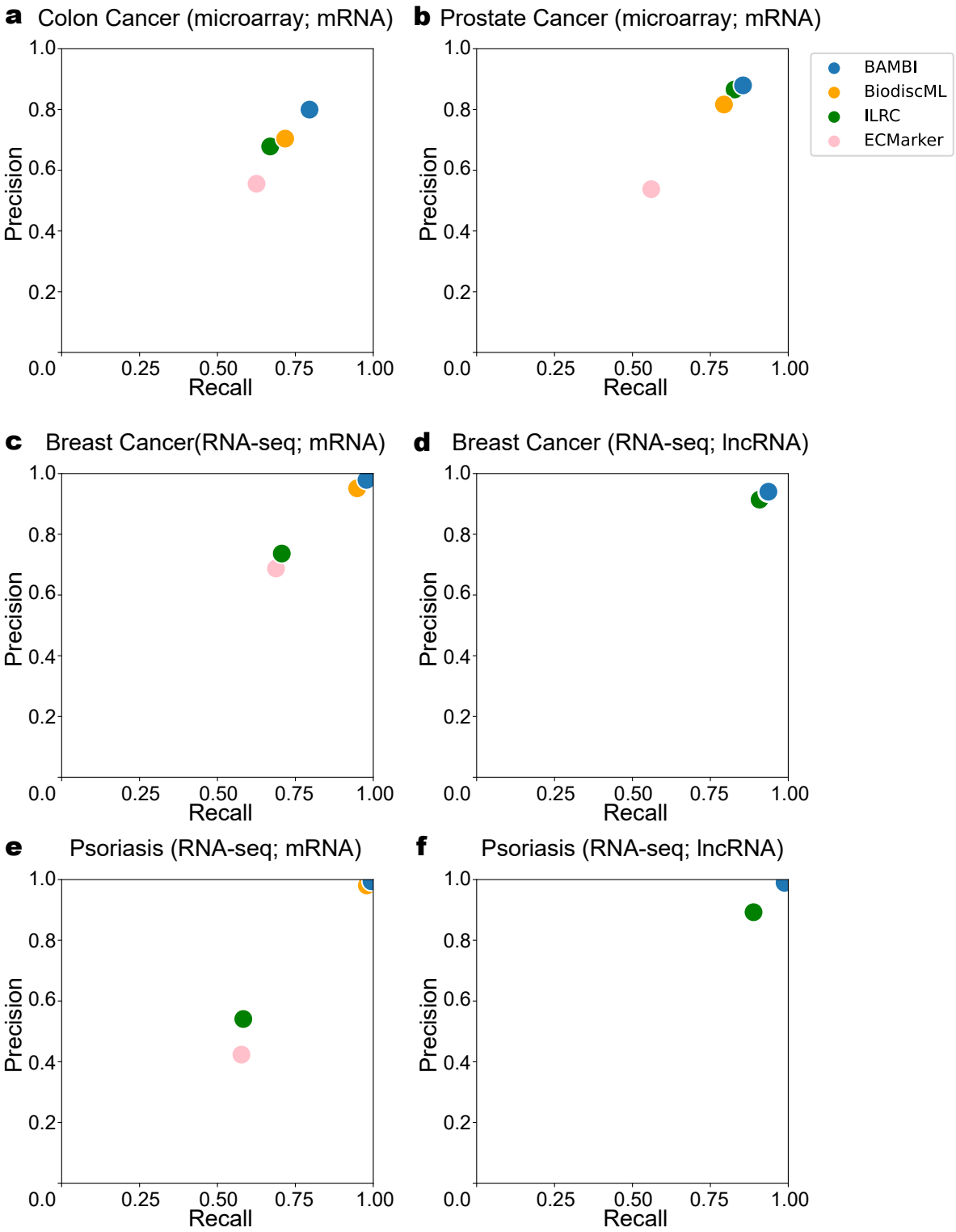
To evaluate the performance of biomarker detection methods under varying sample sizes, we randomly subsampled *n* samples from the simulated datasets, with an equal number of samples from the patient and control groups. The value of n varied across 200, 150, 100, 50, 30, 20, and 10 to simulate scenarios with decreasing sample sizes. Biomarker detection accuracy was assessed for each sample size to compare the method robustness across different sample sizes. This accuracy was calculated as the average proportion of correctly identified target biomarker genes out of the total number of genes identified by the method:

$$Accuracy = \frac{\text{Number of identified targeted}-\text{biomarkers}}{\text{Total number of identified biomarkers}} .$$

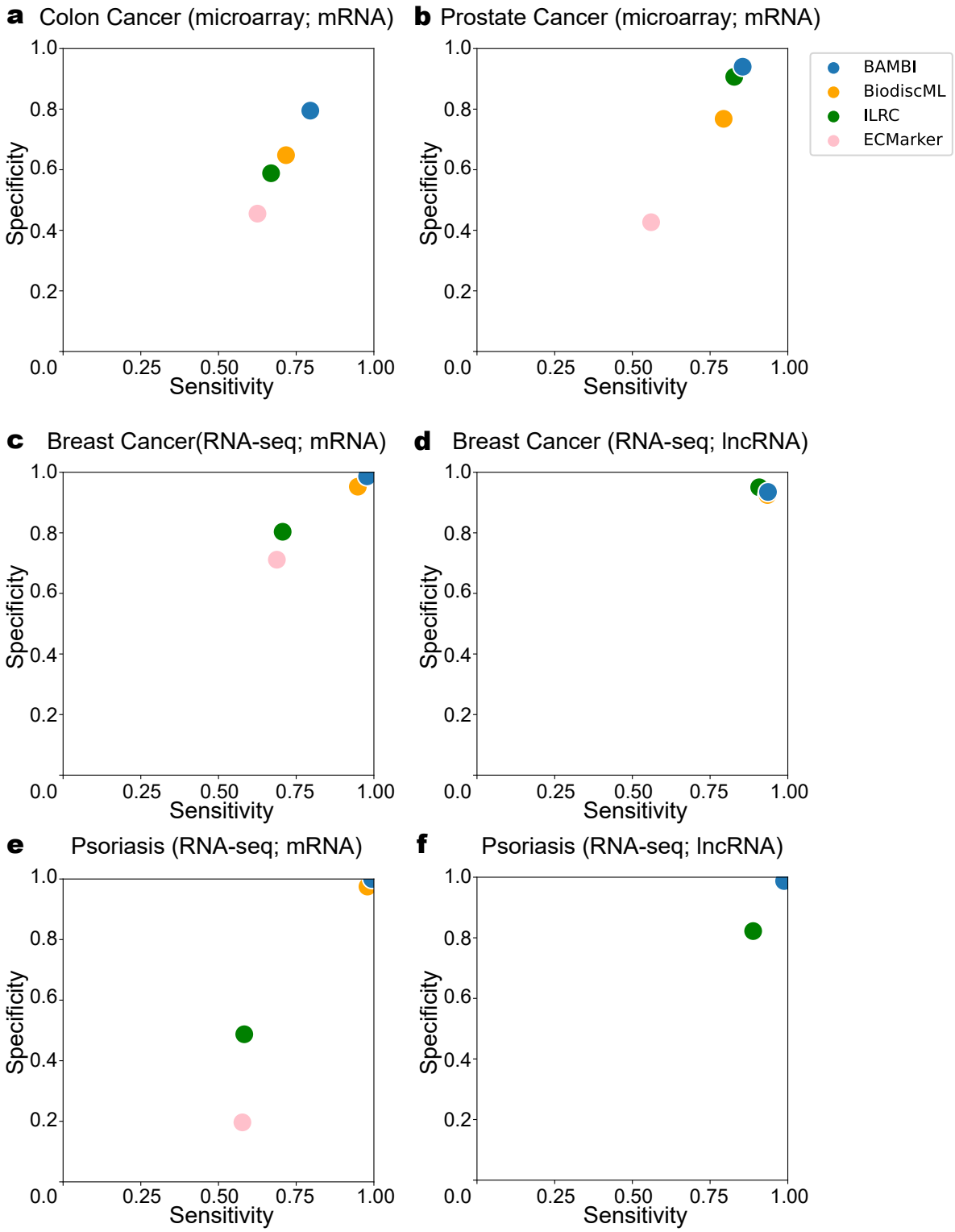
It is important to note that biomarker detection tasks differ fundamentally from traditional prediction tasks and are inherently less prone to overfitting. Overfitting typically occurs in traditional prediction tasks where models use group labels (e.g., patient vs. control) as targeted labels during training. In such scenarios, models may memorize specific patterns, noise, or spurious correlations associated with these labels, leading to overfitting. In contrast, biomarker detection tasks, including those performed by BAMBI and other methods, biomarkers are identified through feature selection and the evaluation of gene importance and relevance to groups. Importantly, in our study, the simulated datasets were designed with pre-defined feature labels (e.g., targeted biomarkers, shuffled biomarkers, and non-biomarkers) that were entirely independent of the model training process. During biomarker detection, models accessed only group labels (e.g., patient vs. control) and corresponding gene expression profiles during biomarker detection. This design inherently prevents models from overfitting to the biomarker labels, as they are not part of the training data.



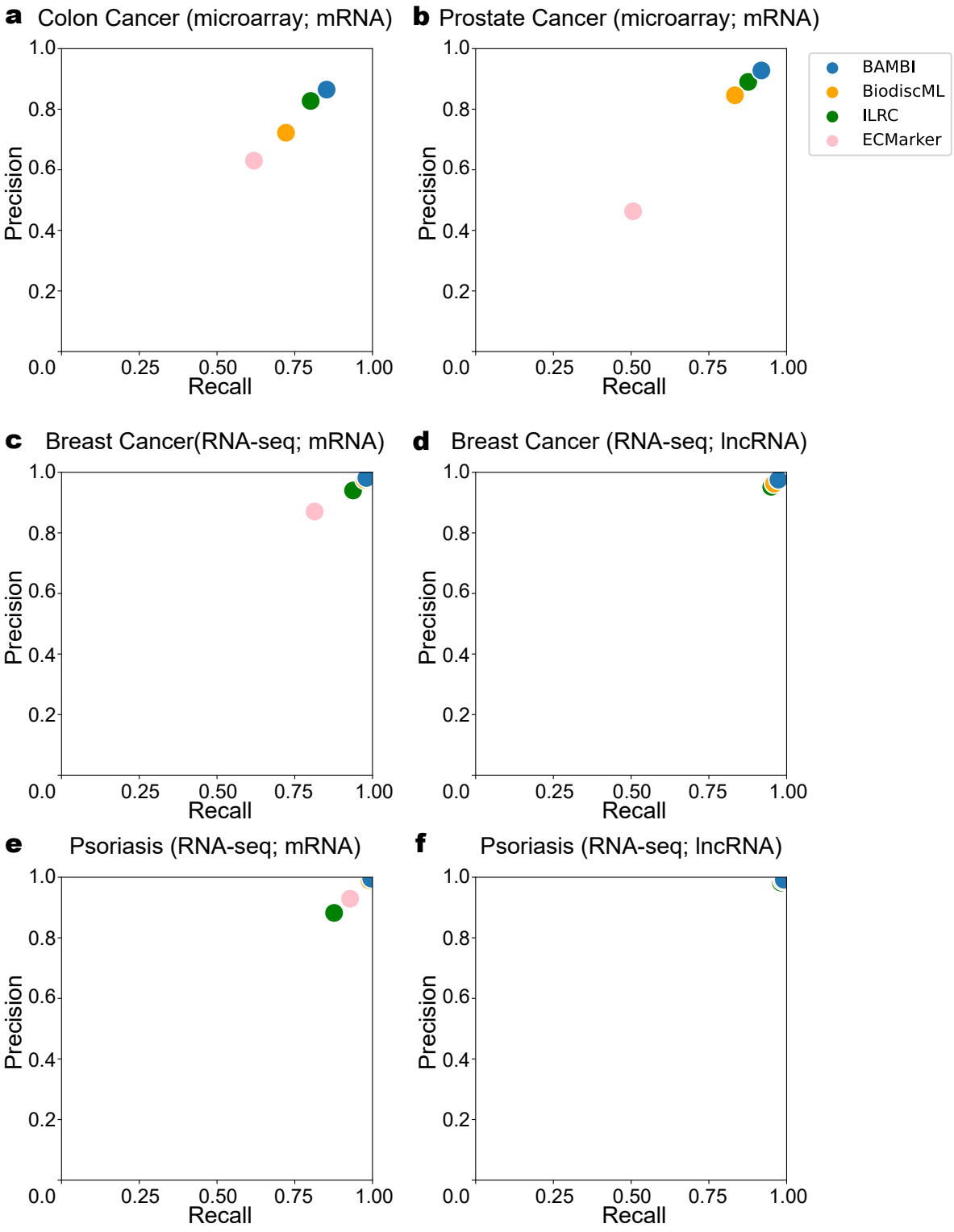
Supplementary Fig. S1: Filtering genes with significant overlapping expression-distributions. In BAMBI, we included the estimated distribution overlapping area of the expression levels between the two groups as a criterion to evaluate genes' expression heterogeneous. Here we list two examples: the expression levels of two lncRNA genes (*HSALNG0082460* and *HSALNG0026574*) in the TCGA breast cancer dataset. (**a-d**) Both their mean difference and median difference of the lncRNA gene *HSALNG0082460* (mean difference: 0.14; median difference: 0.15) are larger than the lncRNA gene *HSALNG0026574* (mean difference: 0.11; median difference: 0.07). However, (**e-f**) the lncRNA *HSALNG0082460* has a significant overlap area (31.2%) of its expression distributions between the control and patient groups, while the lncRNA *HSALNG0026574* has a small overlap area (1.7%) of its expression distribution. The biomarker gene with smaller distribution overlap between groups is more practical for diagnostics, as it allows for clear threshold-based disease prediction, enhancing accuracy and ease of interpretation for clinicians. In this example, *HSALNG0026574* might be more suitable to be used as biomarker than *HSALNG0082460*.



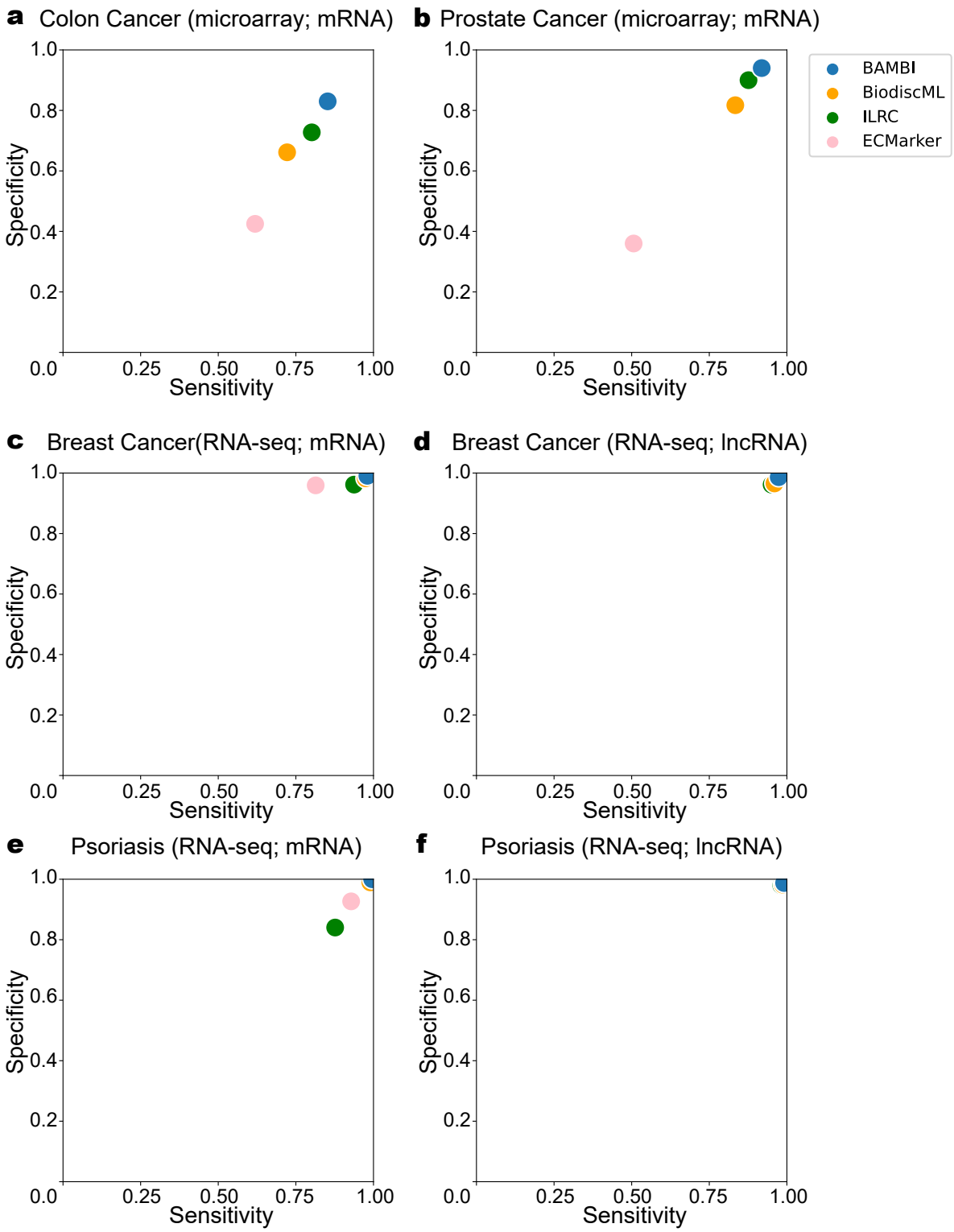
Supplementary Fig. S2: BAMBI surpasses competing methods in achieving higher recall vs. precision rates for identifying single biomarkers. a–f Each scatter plot represents the performance of recall vs. precision for each method in different scenarios.



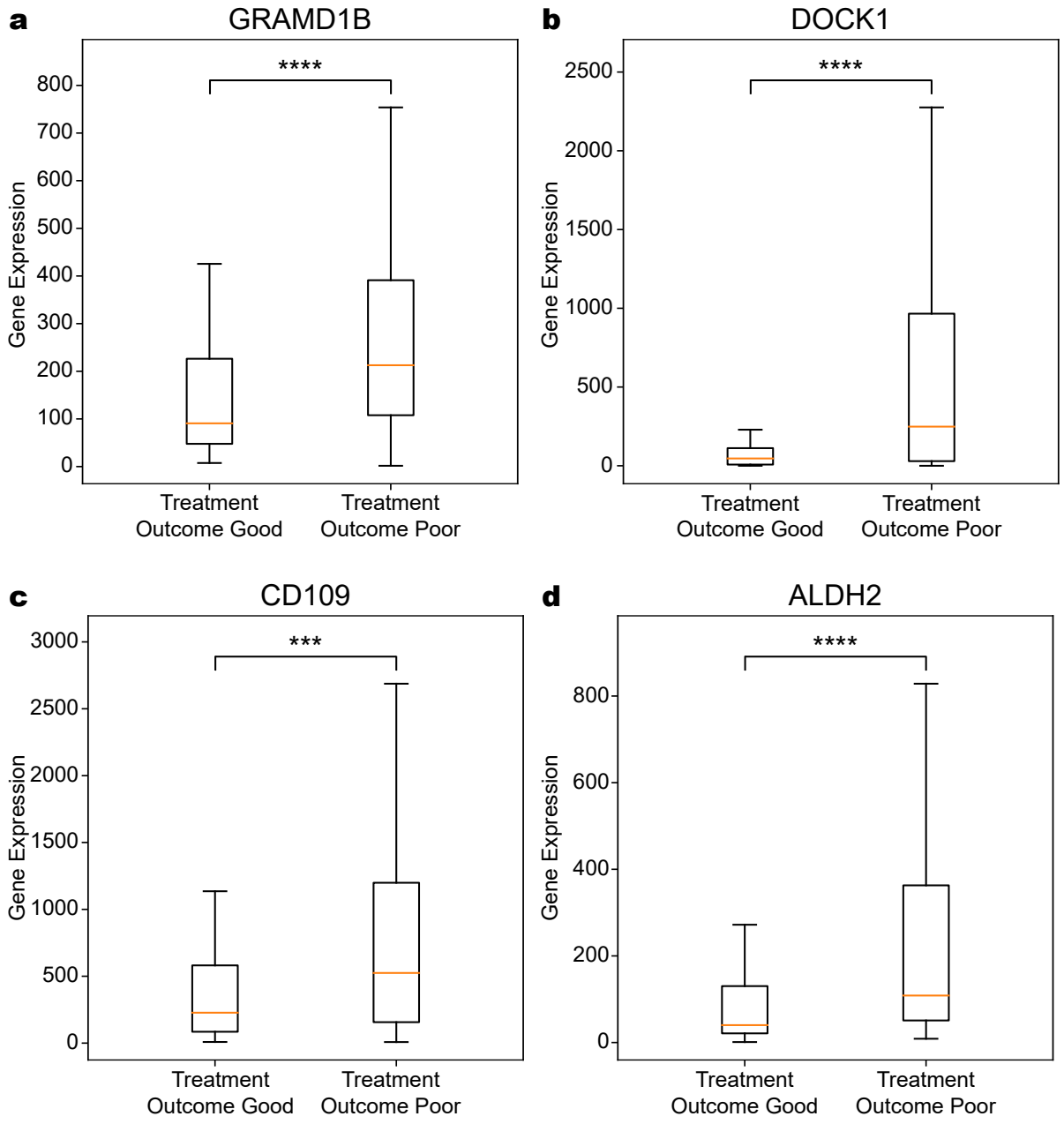
Supplementary Fig. S3: BAMBI surpasses competing methods in achieving higher sensitivity vs. specificity rates for identifying single biomarkers. a–f Each scatter plot represents the performance of sensitivity vs. specificity for each method in different scenarios.



Supplementary Fig. S4: BAMBI surpasses competing methods in achieving higher recall vs. precision rates for identifying panel biomarkers. a-f Each plot represents the performance of recall vs. precision for each method in different scenarios.



Supplementary Fig. S5: BAMBI surpasses competing methods by achieving higher sensitivity vs. specificity rates for identifying panel biomarkers. a–f Each scatter plot represents the performance of sensitivity vs. specificity for each method in different scenarios.



Supplementary Fig. S6: BAMBI identifies putative prognostic biomarkers for acute myeloid leukemia. a-d Four putative prognostic biomarker genes identified by BAMBI are negatively correlated with the treatment outcome in the training cohort. *** indicates $p < 0.001$; **** indicates $p < 0.0001$.

Supplementary Algorithm S1: BAMBI Algorithm Framework in Pseudo-Code

Input: RNA-seq dataset D with n samples and m genes, class labels y (case/control), Initialize gene set $G \leftarrow \{g_1, g_2, \cdots, g_m\}$

Output: single RNA Biomarkers ($G_{singular}$) and panel of multiple RNA Biomarkers (G_{panel})

# Biologically Informed Statistical Feature Selection
for $g_i \in \{g_1, g_2, \cdots, g_m\}$ **do**
Compute p-value using Wilcoxon test for case vs. control groups p_value_i
Compute Fold-change between case and control groups $Fold\_change_i$
Compute maximum expression levels $maximum\_expression_i$
Estimate the expression distributions of case and control groups, compute overlap area $Expression\_overlap\_area_i$
end for
Update $G \rightarrow G^*$: retain genes that satisfy all of the following criteria:
- $p\_value_i \leq threshold_{p\_value}$
- $Fold\_change_i \geq threshold_{fold\_change}$
- $maximum\_expression_i \geq threshold_{maximum\_expression}$
- $Expression\_overlap\_area_i \leq threshold_{expression\_overlap\_area}$

Update $D \rightarrow D^*$based on G^*

Machine learning-based feature selection
Split D^* into 10 folds $fold_1, fold_2, \cdots, fold_{10}$, ensuring each fold maintains the original dataset's class distribution.
for $fold_i \in \{fold_1, fold_2, \cdots, fold_{10}\}$ **do**
Use $fold_i$ as $testing\_set_i$, the rest 9 folds as $training\_set_i$, generate $train\_test\_set\_pair_i$
for machine learning $model\_type_j \in \{model\_type_1, model\_type_2, \cdots, model\_type_p\}$ **do**
Initialize $G_{current} \leftarrow G^*$
while $|G_{current}| > 0$:
Train model $model\_type_j$ on $training\_set_i$ with $G_{current}$
Compute model score $Score_{current}$ with $G_{current}$ by Repeated Stratified K-Fold cross validation
Compute SHAP values for each gene in $G_{current}$
Remove the least prediction influential gene g_{min} based on SHAP scores
end while
Generate a continuous curve plot based on collected $Score_{current}$ against corresponding gene count
Identify knee point of curve, select corresponding gene set as $optimal\_gene\_set_{ij}$
Perform hyperparameter tuning & model training on $model\_type_j$ with $optimal\_gene\_set_{ij}$ by Exhaustive Grid search with cross validation
Evaluate the trained model using $testing\_set_i$ to compute the performance metric $Evaluation\_Score_{ij}$ (e.g., accuracy, AUC, F1-score)
end for
end for

Enhanced ten-fold cross-validation for robust biomarker selection
Construct Evaluation Table
Construct an Evaluation Score table S with dimensions $10 \times p$
- Each row corresponds to a $train\_test\_set\_pair_i$ ($i\ in\ \{1,\ 2, \cdots,\ 10\}$)
- Each column corresponds to a $model\_type_j$ ($j\ in\ \{1,\ 2, \cdots,\ p\}$)
- Cells include $Evaluation\_Score_{ij}$ with corresponded $optimal\_gene\_set_{ij}$

Filter Poor-Performing Models
for $train\_test\_set\_pair_i \in \{train\_test\_set\_pair_1, \cdots, train\_test\_set\_pair_{10}\}$ **do**
Identify $Best\_Evaluation\_Score_i$ as the highest performance score across all models in $train\_test\_set\_pair_i$
Remove models in $train\_test\_set\_pair_i$ with $Evaluation\_Score_{ij} < 95\% \times Best\_Evaluation\_Score_i$ from Evaluation Score table S
end for
Get update Evaluation Score table S^*

Single Biomarker Identification
Aggregate Evaluation Score table S^* into gene level
for each g_k appeared in the $optimal\_gene\_set_{ij}$ of S^* **do**
Compute $Pairs_{g_k}$: Number of train-test pairs where g_k is selected (indicating robustness across splits).
Compute $Model\_types_{g_k}$: Number of model types including g_k (suggesting diversity of selection).
Compute $Frequency_{g_k}$: Total occurrences of g_k across S^* (indicates overall reliability).
end for
Prioritize genes based on considerations: $Pairs_{g_k} > Model\_types_{g_k} > Frequency_{g_k}$
Select top-ranked genes as Single Biomarkers $G_{singular}$

Panel Biomarker Identification
Aggregate Evaluation Score table S^* into model level based on identical $optimal\_gene\_set_{ij}$ and $model\_types_j$
for each $Model_k$ appeared in S^* **do**
Compute $Pairs_{Model_k}$: Number of train-test pairs covered by the $Model_k$ (indicating robustness across splits).
Compute $AvgPerformance_{Model_k}$: Average test set performance (indicating predictive).
Compute $GeneCount_{Model_k}$: Number of genes in $Model_k$.
end for
Prioritize model based on considerations:
- $Pairs_{Model_k} > AvgPerformance_{Model_k} > GeneCount_{Model_k}$
- if all $Pairs_{Modelk} \leq 2$, $AvgPerformance_{Model_k} > Pairs_{Model_k} > GeneCount_{Model_k}$

Select top-ranked models as a panel of multiple RNA Biomarkers G_{panel}

Supplementary Table S1

Dataset	Methods	BalAcc	Sens	Prec	Spec	F1	AUC	Acc
Colon_Cancer_microarray	BAMBI	79.54%	79.54%	79.97%	79.50%	77.07%	79.54%	79.08%
Colon_Cancer_microarray	BiodiscML	71.73%	71.73%	70.41%	64.80%	67.80%	71.73%	73.33%
Colon_Cancer_microarray	ILRC	66.92%	66.92%	67.83%	58.83%	64.84%	66.92%	69.89%
Colon_Cancer_microarray	ECMarker	62.54%	62.54%	55.55%	45.50%	56.54%	62.54%	67.54%
Prostate_Cancer_microarray	BAMBI	85.48%	85.48%	87.95%	94.00%	84.99%	85.48%	85.43%
Prostate_Cancer_microarray	BiodiscML	79.36%	79.36%	81.64%	76.80%	78.48%	79.36%	79.44%
Prostate_Cancer_microarray	ILRC	82.77%	82.77%	86.60%	90.67%	81.42%	82.77%	82.82%
Prostate_Cancer_microarray	ECMarker	56.03%	56.03%	53.76%	42.67%	49.49%	56.03%	56.21%
TCGA_Breast_Cancer_PC	BAMBI	97.82%	97.82%	97.91%	98.66%	97.80%	97.82%	97.81%
TCGA_Breast_Cancer_PC	BiodiscML	94.83%	94.83%	95.14%	95.25%	94.81%	94.83%	94.83%
TCGA_Breast_Cancer_PC	ILRC	70.64%	70.64%	73.66%	80.36%	67.91%	70.64%	70.57%
TCGA_Breast_Cancer_PC	ECMarker	68.78%	68.78%	68.72%	71.13%	63.30%	68.78%	68.50%
TCGA_Breast_Cancer_lncRNA	BAMBI	93.62%	93.62%	94.03%	93.50%	93.62%	93.62%	93.64%
TCGA_Breast_Cancer_lncRNA	BiodiscML	93.46%	93.46%	93.77%	92.49%	93.45%	93.46%	93.47%
TCGA_Breast_Cancer_lncRNA	ILRC	90.80%	90.80%	91.41%	95.04%	90.64%	90.80%	90.73%
GSE54456_PC	BAMBI	99.44%	99.44%	99.41%	100.00%	99.41%	99.44%	99.41%
GSE54456_PC	BiodiscML	97.93%	97.93%	98.04%	97.50%	97.93%	97.93%	97.96%
GSE54456_PC	ILRC	58.29%	58.29%	54.07%	48.74%	53.26%	58.29%	59.10%
GSE54456_PC	ECMarker	57.69%	57.69%	42.34%	19.64%	46.29%	57.69%	59.78%
GSE54456_lncRNA	BAMBI	98.82%	98.82%	98.93%	98.75%	98.84%	98.82%	98.85%
GSE54456_lncRNA	BiodiscML	99.10%	99.10%	99.21%	98.75%	99.13%	99.10%	99.13%
GSE54456_lncRNA	ILRC	88.86%	88.86%	89.25%	82.22%	87.44%	88.86%	89.22%
Average	BAMBI	92.45%	92.45%	93.03%	94.07%	91.96%	92.45%	92.37%
Average	BiodiscML	89.40%	89.40%	89.70%	87.60%	88.60%	89.40%	89.69%
Average	ILRC	76.38%	76.38%	77.14%	75.98%	74.25%	76.38%	77.05%
Average	ECMarker	61.26%	61.26%	55.09%	44.73%	53.91%	61.26%	63.01%

Supplementary Table S1: Detailed Single Biomarker performance metrics for BAMBI and the other three methods across various datasets. Metrics include balanced accuracy(BalAcc), sensitivity(Sens), precision(Prec), specificity(Spec), F1 score(F1), Area under the curve(AUC) and accuracy(Acc), offering a comprehensive comparison of their capabilities in identifying RNA biomarkers.

Supplementary Table S2

Dataset	Methods	BalAcc	Sens	Prec	Spec	F1	AUC	Acc
Colon_Cancer_microarray	BAMBI	85.25%	85.25%	86.47%	83.00%	84.46%	85.25%	85.64%
Colon_Cancer_microarray	BiodiscML	72.18%	72.18%	72.23%	66.15%	71.02%	72.18%	73.71%
Colon_Cancer_microarray	ILRC	80.12%	80.12%	82.73%	72.75%	79.33%	80.12%	82.02%
Colon_Cancer_microarray	ECMarker	61.88%	61.88%	63.01%	42.50%	58.92%	61.88%	67.88%
Prostate_Cancer_microarray	BAMBI	91.89%	91.89%	92.80%	94.00%	91.87%	91.89%	91.94%
Prostate_Cancer_microarray	BiodiscML	83.38%	83.38%	84.60%	81.72%	83.19%	83.38%	83.41%
Prostate_Cancer_microarray	ILRC	87.64%	87.64%	89.01%	90.00%	87.53%	87.64%	87.71%
Prostate_Cancer_microarray	ECMarker	50.64%	50.64%	46.30%	36.00%	46.54%	50.64%	51.00%
TCGA_Breast_Cancer_PC	BAMBI	98.03%	98.03%	98.13%	99.09%	98.03%	98.03%	98.03%
TCGA_Breast_Cancer_PC	BiodiscML	97.47%	97.47%	97.57%	98.26%	97.45%	97.47%	97.46%
TCGA_Breast_Cancer_PC	ILRC	93.76%	93.76%	94.00%	96.18%	93.68%	93.76%	93.71%
TCGA_Breast_Cancer_PC	ECMarker	81.39%	81.39%	87.05%	95.93%	78.40%	81.39%	81.04%
TCGA_Breast_Cancer_lncRNA	BAMBI	97.36%	97.36%	97.58%	98.64%	97.36%	97.36%	97.37%
TCGA_Breast_Cancer_lncRNA	BiodiscML	95.99%	95.99%	96.27%	96.56%	95.98%	95.99%	96.00%
TCGA_Breast_Cancer_lncRNA	ILRC	95.07%	95.07%	95.21%	96.19%	95.06%	95.07%	95.06%
GSE54456_PC	BAMBI	99.58%	99.58%	99.56%	100.00%	99.56%	99.58%	99.56%
GSE54456_PC	BiodiscML	98.99%	98.99%	99.03%	98.97%	98.99%	98.99%	99.00%
GSE54456_PC	ILRC	87.66%	87.66%	88.22%	84.03%	87.54%	87.66%	87.82%
GSE54456_PC	ECMarker	92.80%	92.80%	92.91%	92.68%	92.78%	92.80%	92.83%
GSE54456_lncRNA	BAMBI	99.10%	99.10%	99.21%	98.75%	99.13%	99.10%	99.13%
GSE54456_lncRNA	BiodiscML	98.71%	98.71%	98.85%	98.22%	98.74%	98.71%	98.75%
GSE54456_lncRNA	ILRC	98.17%	98.17%	98.23%	98.01%	98.18%	98.17%	98.18%
Average	BAMBI	95.20%	95.20%	95.62%	95.58%	95.07%	95.20%	95.28%
Average	BiodiscML	91.12%	91.12%	91.42%	89.98%	90.90%	91.12%	91.39%
Average	ILRC	90.40%	90.40%	91.23%	89.53%	90.22%	90.40%	90.75%
Average	ECMarker	71.68%	71.68%	72.32%	66.78%	69.16%	71.68%	73.19%

Supplementary Table S2: Detailed Panel Biomarkers performance metrics for BAMBI and the other three methods across various datasets. Metrics include balanced accuracy(BalAcc), sensitivity(Sens), precision(Prec), specificity(Spec), F1 score(F1), Area under the curve(AUC) and accuracy(Acc), offering a comprehensive comparison of their capabilities in identifying RNA biomarkers.

Supplementary Table S3

Number of Samples	BAMBI	BioDiscML	ILRC	ECMarker
10	63.82%	42.61%	31.00%	0.00%
20	100.00%	79.94%	57.00%	1.50%
30	100.00%	86.57%	64.50%	1.00%
50	100.00%	89.80%	56.50%	1.50%
100	100.00%	82.07%	58.00%	5.50%
150	100.00%	84.70%	94.00%	7.50%
200	100.00%	89.47%	99.50%	8.00%

Supplementary Table S3: Detailed biomarker detection accuracy for BAMBI and the other three methods across varying sample sizes in the simulated datasets. This table complements Fig. 2 g by providing the exact accuracy values for each method at sample sizes ranging from 10 to 200, highlighting BAMBI's robustness and ability to handle small sample sizes under different data availability scenarios.

Reference

1. Li Y, Ge X, Peng F, et al. Exaggerated false positives by popular differential expression methods when analyzing human population samples. Genome Biol 2022; 23:
2. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014; 15:
3. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2009; 26:139–140
4. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43:e47
5. Sigova AA, Mullen AC, Molinie B, et al. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc Natl Acad Sci U S A 2013; 110:2876–2881
6. 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:
7. Li Z, Zhou P, Kwon E, et al. Flnc: Machine Learning Improves the Identification of Novel Long Noncoding RNAs from Stand-Alone RNA-Seq Data. Noncoding RNA 2022; 8:
8. Lundberg SM, Allen PG, Lee S-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS) 2017; 30:4765–4774
9. Hasan MM, Basith S, Khatun MS, et al. Meta-i6mA: An interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. Brief Bioinform 2021; 22:
10. Hasan MM, Schaduangrat N, Basith S, et al. HLPpred-Fuse: Improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. Bioinformatics 2020; 36:3350–3356
11. Hasan MM, Alam MA, Shoombuatong W, et al. NeuroPred-FRL: An interpretable prediction model for identifying neuropeptide using feature representation learning. Brief Bioinform 2021; 22:
12. 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:
13. 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:
14. 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; 37:1115–1124
15. Weinstein JN, Collisson EA, Mills GB, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet 2013; 45:1113–1120
16. 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. Journal of Investigative Dermatology 2014; 134:1828–1838
17. Singh D, Febbo PG, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 2002; 1:203–209
18. 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 U S A 1999; 96:6745–6750