A robust and generalizable immune-related signature for sepsis diagnostics

Yueran Yang[1,2,#], Yu Zhang[1,2,#], Shuai Li[2], Xubin Zheng[1,3], Man-Hon Wong[3], Kwong-Sak Leung[3], Lixin Cheng[1,*]

Abstract—High-throughput sequencing can detect tens of thousands of genes in parallel, providing opportunities for improving the diagnostic accuracy of multiple diseases including sepsis, which is an aggressive inflammatory response to infection that can cause organ failure and death. Early screening of sepsis is essential in clinic, but no effective diagnostic biomarkers are available yet. Here, we present a novel method, Recurrent Logistic Regression, to identify diagnostic biomarkers for sepsis from the blood transcriptome data. A panel including five immune-related genes, LRRN3, IL2RB, FCER1A, TLR5, and S100A12, are determined as diagnostic biomarkers (LIFTS) for sepsis. LIFTS discriminates patients with sepsis from normal controls in high accuracy (AUROC = 0.9959 on average; IC = [0.9722-1.0]) on nine validation cohorts across three independent platforms, which outperforms existing markers. Our analysis determined an accurate prediction model and reproducible transcriptome biomarkers that can lay a foundation for clinical diagnostic tests and biological mechanistic studies.

Index Terms—sepsis, transcriptome, signature, immune genes, diagnosis

✦

1 INTRODUCTION

SEPSIS is a life-threatening organ dysfunction caused by a host's unbalanced response to an infection. It is one of the most severe diseases in the intensive care unit (ICU) and one of the world's leading lethal diseases [1] [2] [3] [4]. Its common clinical manifestations include abnormalities in body temperature, heart rate, breathing, and peripheral white blood cell counts. Besides, sepsis is often accompanied by multiple organ dysfunction syndromes, such as hemodynamic instability, respiratory failure, and disseminated intravascular coagulation. In the past few decades, the high morbidity and mortality caused by sepsis have made the society to endure huge economic burden [1] [2] [3]. The prevalent methods of the diagnosis of sepsis are microbiological culture and taxonomic identification of the pathogen. However, the methods based on bacterial culture techniques have several shortcomings: (1) it usually takes 24 hours to obtain a positive result; (2) only one-third of the blood cultures are positive in clinically diagnosed sepsis cases, so negative results in culture do not mean negative cases; (3) the chances of a positive culture are reduced in patients who have already used antibiotics; (4) false positives are frequently caused by contamination; (5) short-term bacteremia can lead to a positive blood culture without a severe inflammatory response. Therefore, the sensitivity and

specificity of the methods based on microbiological culture are quite low, and hence fails to diagnose sepsis effectively. [5] [6]

Biomarkers such as procalcitonin (PCT) and C-reactive protein (CRP) have been considered to diagnose and evaluate sepsis in emergency department (ED) and intensive care unit (ICU). PCT is increasingly recognized as an important diagnostic and monitoring tool in clinical practice that provides significant information for clinical decision making. It is a potential biomarker in assisting clinicians in the diagnosis of generalized infection and sepsis in ICU. Several systematic reviews and meta-analyses have been carried out to describe the utility of PCT in distinguishing sepsis from SIRS and non-septic burn patients. [7] However, the overall sensitivity and specificity of PCT range from 0.72 to 0.93 and 0.64 to 0.84, respectively [8], which is incompetent in the clinical context. CRP was reported as an indicator whose daily measurement is useful in monitoring sepsis, but its low specificity may be its primary drawback and it is hard to define CRP as an independent predictor of sepsis [9].

In recent years, with the rise of high-throughput sequencing technology, tens of thousands of genes can be detected in parallel [10] [11] [12] [13], providing opportunities for precise molecular diagnosis using machine learning methods [14] [15] [16] [17] [18]. Several gene markers have been developed for the diagnostic prediction of sepsis. For instance, Scicluna *et al.* proposed the FAIM3:PLAC8 ratio as a candidate biomarker to assist in the rapid diagnosis of community-acquired pneumonia (CAP) [19], which accounts for a high proportion of intensive care unit (ICU) admissions for respiratory failure and sepsis. McHugh *et al.* designed a classifier SeptiCyte Lab composed of four mRNAs of CEACAM4, LAMP1, PLA2G7, and PLAC8 by applying Support Vector Machine (SVM) and Random Forest [20]. Scicluna *et al.* developed a sNIP score for sepsis

1. *Shenzhen People's Hospital, First Affiliated Hospital of Southern University of Science and Technology, The Second Clinical Medicine College of Jinan University, Shenzhen, China*
2. *John Hopcroft Center for Computer Science, Shanghai Jiao Tong University, Shanghai, China*
3. *Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China*
\# *These authors contribute equally to this study.*
* *Corresponding author: Lixin Cheng, Ph.D., Shenzhen People's Hospital, 1017 Dongmen North Road, Shenzhen, Guangdong Province 518020, China; Email: easonlcheng@gmail.com*

diagnosis based on the expression abundance of three genes using SVM and bootstrapping [21]. However, the above-mentioned mRNA biomarkers cannot obtain consistent results in multiple independent data sets.

In this paper, we introduced a novel recurrent logistic regression (RLR) as an automatic detection for the diagnostic biomarkers of sepsis. Since patients with sepsis have a severely dysregulated immune system [20] [21] [22], we principally concentrated on the immune-related genes (IRGs) and regarded them as the key molecular events involved in sepsis. Based on IRGs, the RLR model was trained and the less significant genes were filtered during each iteration until no gene is eliminated. Regularization and elimination of insignificant features were applied simultaneously in RLR to avoid overfitting and hence reduce the complexity of the discriminative model. The biomarkers identified by RLR were verified on nine independent expression cohorts across three different detection platforms. We also evaluated the classification performance of each individual gene in the identified biomarkers. Finally, network and functional analyses were carried out for the genes interacting with these biomarkers.

2 MATERIALS AND METHODS

2.1 Data and preprocessing

Eleven different gene expression cohorts were collected from the Gene Expression Omnibus (GEO) database with both sepsis samples and healthy controls, including three adult datasets and eight pediatric datasets (Table 1). In total, 1,384 samples were analyzed from three microarray platforms, Affymetrix Human Genome U133 Plus 2.0 (AffyU133P2), Affymetrix Human Genome U219 (AffyU219), and Agilent Human Gene Expression 4x44K v2 Microarray (AgilentV2). The raw data were preprocessed and normalized using the robust multichip average (RMA) algorithm [23] [24] [25].

GSE57065 is adopted for model training (Discovery cohort I) and GSE26378 is used for tuning the hyperparameter (Discovery cohort II). Seven datasets (GSE95233, GSE28750, GSE8121, GSE13904, GSE26440, GSE9692, and GSE4607) from AffyU133P2 serve as the validation cohorts I to evaluate the diagnostic performance. GSE65682 and E-MTAB-1548 detected by other platforms are set as the validation cohorts II for evaluating the cross-platform capability.

We intend to train a robust prediction model across the biological heterogeneity of childhood and adult sepsis, so the adult and children samples were incorporated for model training.

2.2 Immune-related gene selection

Since sepsis is a disease related to patients' immune systems, only immune-related genes (IRGs) are considered as potential biomarkers in this study. 770 IRGs were collected from the database nanoString (www.nanoString.com), which has been used in hundreds of studies of pathogen infection and the related host response. [26] [27] The numbers of IRGs are 737, 740, and 627 on AffyU133P2, AffyU219, and AgilentV2, respectively. We aimed to find a biomarker that can be applied to different platforms, so the 608 common

IRGs of the three platforms were utilized for computational modeling (Figure 1a).

2.3 Recurrent logistic regression

Recurrent logistic regression contains many iterations with model optimization and automatical feature selection, since each iteration involves regression step and elimination step (Figure 1b).

Regression step: Logistic regression is employed to the candidate gene set (initially 608 IRGs). The expression abundance of genes in each sample, is represented by a vector denoted as

$$\boldsymbol{x} = (g_1, \cdots, g_n)^T \qquad (1)$$

where g_i is the i-th gene expression.

To construct a classifier involving fewer genes features based on the expression vector $\boldsymbol{x}$ of a sample, a function $f : \mathbb{R}^n \to \{0, 1\}$ is built, where 0 represents healthy controls and 1 represents sepsis. The logistic model applied in RLR is a binary classifier expressed by

$$f(x) = \frac{1}{1 + e^{-\boldsymbol{w}\boldsymbol{x}}} = \frac{1}{1 + e^{-(w_0 + w_1 g_1 + \cdots + w_n g_n)}} \qquad (2)$$

where $\boldsymbol{x}$ is an expression vector and $f(x)$ is a diagnostic risk score used to predict the probability of having sepsis. $\boldsymbol{w} = (w_0, w_1, \cdots, w_n)^T$ are parameters optimized by the cost function with regularization,

$$J(w; X) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i(\boldsymbol{x}_i^T \boldsymbol{w} + w_0)\right)\right) \qquad (3)$$

where X is the collection of samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ in discovery cohort and $y \in \{0, 1\}$ is the label for each sample.

Elimination step: After optimizing the regression model, the minor genes regarded as less significant are eliminated. A gene g_i is defined as minor gene if the absolute value of its corresponding weight w_i is less than a proportion of the absolute maximum weight, i.e.,

$$|w_i| < C \max(|w_1|, \cdots, |w_n|) \qquad (4)$$

where $C \in [0, 1]$ is a hyperparameter. Instead of using the traditional way that chooses a fixed threshold such as P value < 0.01, this step selects features adaptively based on the maximum weight trained by the model.

The regression step and elimination step are repeated iteratively until it converges, specifically, no more minor gene remained. In this sense, the algorithm is named the recurrent logistic regression (RLR).

The RLR is first trained on the discovery cohort I and evaluated by the AUROC on discovery cohort II (Figure 1c). We exhaustedly tried possible values of the hyperparameter C with the search space between 0.75 and 0.9 and each interval equaling to 0.01. The hyperparameter C can therefore be determined by the optimal performance on discovery cohort II.

2.4 Performance evaluation and analysis

Receiver operating characteristic (ROC) curve was applied for performance evaluation, which is the function image of True Positive Rate (TPR) with respect to False Positive Rate

(FPR), where TPR represents the positive correctly classified samples to the total number of positive samples and FPR represents the ratio between the incorrectly classified negative samples to the total number of negative samples. Area Under the Curve (AUC) means the area under the ROC curve ranging from 0 to 1. Higher AUC indicates a more discriminative model. We use AUC to quantify the discrimination ability of the models on seven cohorts measured with the same platform and compare to the existing biomarkers. Moreover, the cross-platform capability is also evaluated on two cohorts from different platforms.

Meta-analysis was conducted for the constructed gene panel LIFTS (LRRN3, IL2RB, FCER1A, TLR5, and S100A12) and the Standardized Mean Difference (SMD) are demonstrated in forest plot (Figure 3). Four graphical elements are presented including the estimated SMD (solid block), the respective 95% confidence intervals for each cohort (horizontal line), the non-effect size (vertical line), and overall estimation of all cohorts (red rhombus) [28].

To analyze the role of the five genes in LFTS, we presented the human protein interaction and constructed the protein interaction network. Protein interactions were obtained from the database InWeb_InBioMap [29], [30]which is the most comprehensive resource for human protein interactome. Around 57% of the interactions are experimentally validated. The interaction network was conducted and illustrated using the R package igraph.Function enrichment was carried out using the R package clusterprofiler [31] and the network-guided gene set characterization pipeline of KownEnG [32], respectively.

3 RESULTS

3.1 Identification of LIFTS

Patients with sepsis have a severely dysregulated immune system, which impairs clearance of the infection and leaves the body susceptible to new infections with an increased risk of death. Thus, we principally concentrated on the immune-related genes (IRGs) and regarded them as the key molecular events involved in sepsis. After taking the intersection across different platforms, 608 IRGs are screened as potential biomarkers for further analysis.

The recurrent logistic regression (RLR) was then applied on the discovery cohort GSE57065. Different hyperparameter results in multiple gene panels. To select the best gene panel and the hyperparameter, we tried a series of thresholds and evaluated their performance on the independent discovery cohort II, GSE26378. Figure 2 displays the AUROC of these gene panels and indicates that generally a larger C results in a smaller model size N during the optimization. We finalized the model with the highest AUROC up to 0.9951 when C equals to 0.83. The discriminative function of the diagnostic model is

$$
\begin{aligned}
y(x) = [1 + \exp(&-0.9305 g_{LRRN3} - 0.9692 g_{IL2RB} \\
&- 0.7378 g_{FCER1A} + 0.8460 g_{TLR5} \\
&+ 0.8905 g_{S100A12} - 0.0153)]^{-1}
\end{aligned}
\tag{5}
$$

which contains five genes, LRRN3, IL2RB, FCER1A, TLR5, and S100A12. We abbreviated the biomarkers by LIFTS, which is composed by the initial letters of each gene. The

Genome characteristics of the five genes are listed in Table 2.

3.2 The diagnostic capability

Since the value of logistic model is too concentrated to illustrate, i.e., ranging from 0 to 1, we used the corresponsive part in logits of our diagnostic model to represent the diagnostic ability of each gene and LIFTS. Specifically, the logit is

$$
\begin{aligned}
logit = &-0.9305 g_{LRRN3} - 0.9692 g_{IL2RB} \\
&- 0.7378 g_{FCER1A} + 0.8460 g_{TLR5} \\
&+ 0.8905 g_{S100A12}
\end{aligned}
\tag{6}
$$

The standard difference mean $\bar{X} - \bar{Y}$ in effect size between the sepsis and control subjects is displayed in Figure 3, where $\bar{X}$ is the mean of logits for the sepsis samples and $\bar{Y}$ corresponds to normal samples. The five genes individually are qualified to distinguish sepsis samples with the average standard mean difference (SMD) ranging from 1.5 to 3.5 and their confidence intervals do not cross zero. Compared with the individual genes, LIFTS achieves a much higher average SMD of 11.6, suggesting that the weighted gene panel has better diagnostic capability than each of the five genes.

3.3 Performance comparison across different models

LIFTS was evaluated on the nine independent cohorts and compared to existing biomarkers. Figure 4 shows the ROC curves comparison between the LIFTS and three known transcriptome biomarkers, i.e., FAIM3:PLAC8, SeptiCyte Lab, and sNIP. SeptiCyte Lab includes four genes and its risk score is PLAC8/PLA2G7*LAMP1/CEACAM4. sNIP contains three genes and it is represented as (NLRP1-IDNK)/PLAC8. The genes of all these four biomarkers are detectable on the AffyU133P2.

Overall, LIFTS outperforms the other biomarkers on all the validation cohorts except GSE95233. The area under ROC curve (AUC) of LIFTS on each dataset is consistently close to 1. The lowest score, 0.9722 on GSE13904, is still much higher than the other biomarkers. NLRP1 and PLAC8 are not detected on either the AffyU219 or the Agilent platform, so sNIP cannot be applied on dataset GSE65682. Since NLRP1 does not exist on GSE65682 and PLAC8 is not available on E-MTAB-1548, some previous biomarkers could not be evaluated on these two datasets. The AUC of LIFTS on GSE65682 and E-MTAB-1548 are 0.9994 and 1.0, which are superior to its counterparts, indicating the portability of LIFTS among different platforms in diagnostic prediction (Figure 5).

The AUC curves are related to the standard difference means. Considering LIFTS performance shown in Figure 3 and Figure 4, the higher AUC value always corresponds to higher standard difference mean. For example, in GSE13904, the AUC value is relatively low and correspondingly the standard difference mean is relatively closer to zero. However, focusing on E-MTAB-1548, the AUC is 1.0 and the standard difference mean is far from zero.

3.4 Topological and functional analysis of LIFTS

Proteins usually group together as modules to implement in particular cellular functions through interactions [33] [34] [35] [36]. The interference in protein interactions and new undesired protein interactions can cause diseases [37] [38] [39].To explore the functions of the genes in LIFTS, we further studied the topological property of the genes physically interacted with the five genes by network analysis (Figure 6A). Specifically, 1, 45, 19, 7, and 3 partner genes interact with LRRN3, IL2RB, FCER1A, TLR5, and S100A12, respectively (Figure 6B). These genes are closely connected and involved in specific biological processes, including growth hormone synthesis, secretion and action, chemokine signaling pathway, B cell receptor signaling pathway, T cell receptor signaling pathway, *etc.* (Figure 6C). For the protein interaction network, the connections among genes are significantly dense than the simulated networks (P < e-26, Rank sum test, Figure 6D), where we randomly picked up the same number of proteins 10,000 times and calculated their network density distribution. The densities of the simulated networks are mainly less than 0.05 whereas the density of the curated network is 0.2854, indicating the five genes tend to function together with higher network connectivity than other genes.

Given that there are only five genes in LIFTS, standard methods for enrichment analysis may not detect any relevant functional category or pathway. We also employed the network-guided gene set characterization pipeline of KnowEnG [32] for this gene set to better understand their function. Four function resources were used in this pipeline, including Gene Ontology, Enrichr, Pathway Comments, and Reactome. In addition to the functions the partner genes enriched, the five genes are also implemented in pathways of immune system, IL1 and megakaryotyces in obesity, *etc.* Default parameters were used during the analysis.

4 DISCUSSION

We developed a novel model to screen the diagnostic biomarkers of sepsis based on the logistic regression. Five genes were identified as a prediction model (LIFTS) with an average AUROC of 0.9959 among 11 cohorts containing in total 1,384 samples. LIFTS demonstrated its robust portability across three different transcriptome platforms, which is much better than its counterparts such as SeptiCyte Lab [21]. Our analysis thus determined an accurate prediction model and reproducible transcriptome biomarkers that can lay a foundation for clinical diagnostic tests and biological mechanistic studies.

The model was built starting with the immune-related genes. The expression of immune-related genes is response for the dysregulated host immune system to infection in sepsis, so the immune-related genes serve as prior knowledge for the classification model and prevent overfitting, resulting in a robust model for patient heterogeneity. Otherwise, if start with all genes, a different gene panel will be obtained with unexpected noise. The model may get an extremely high performance for the training cohort, but performs worse when it comes to the validation cohorts.

After filtering the genes, the number of candidate genes was greatly shortlisted, which is also an efficient preprocessing step for feature selection. Some researchers made use of differentially expressed (DE) genes as diagnostic signatures [40] [41]. However, DE genes are extremely inconsistent among different datasets and platforms. Only a few overlapping DE genes were obtained among the 11 datasets we used (Supplementary Figure S2, S3), leading to obstacles to find a robust classifier based on DE genes.

The classical logistic regression can only return the classifier with a given number of genes. Mathematically, our goal is to maximize the performance of classification and minimize the complexity of the diagnostic model simultaneously, requiring a competent algorithm with the ability to filter out irrelevant genes automatically. To this end, an enhanced version of logistic regression, recurrent logistic regression (RLR), was developed using the weight of each term as a measure of the gene importance. Importantly, the selection of features and the construction of classifiers are usually regarded as two independent tasks, but we combined these two tasks together. Thus, the biomarkers are more adaptive to the base model and superior to other models, which use a specific algorithm on previously selected biomarkers. When compared with least absolute shrinkage and selection operator (LASSO), a method that useS L1 regularization to impose sparsity, our results demonstrated that RLR overall outperforms LASSO according to AUROC in the discovery and validation cohorts (Supplementary Figure S1).

During the training process, interestingly, we observed that running logistic regression using different coding languages may lead to different results. In this study, the function LogisticRegression in the sklearn package of python was used. Technically, we applied L_2 regularization in our logistic regression processes, which is commonly used in machine learning to reduce model overfitting. In some studies, logistic regression with L_2 regularization is called ridge regression [42]. The advantage of regularization is that it improves numerical stability, not only forces weights to shrink but also copes with the case sophisticatedly when the number of features is larger than the number of samples.

Despite 11 public datasets were taken advantage, all of them were detected using microarray technology. No RNA-seq datasets were included, thereby making our model not applicable for all transcriptome platforms. Therefore, we call for the detection of sepsis using RNA-seq technologies in the near future. Then a large-scale of datasets will be available for further validation, which is able to reduce the risk of the diagnostic model.

Moreover, the proposed method can be used in biomarker identification of other diseases. Since logistic regression is widely used for biomarker identification and several such types of gene expression signatures have been developed for cancers with decent performance, RLR is an upgrade of logistic regression hence it can be applied to the domains where logistic regression can be applied. In terms of the performance, it is superior to logistic regression theoretically, but in practice it depends on a series of factors, such as the detected feature numbers, the sample size, and disease heterogeneity.

In conclusion, the diagnostic biomarkers LIFTS shows higher accuracy and robustness compared to the existing biomarkers when differentiating the sepsis patients from the

normal controls. Further clinical trials are needed to confirm the findings in the paper.

DECLARATIONS

Ethics approval and consent to participate

Not appliable.

Consent for publication

Not appliable.

Competing interests

None decleared.

Authors' contributions

L.C. conceived the project and wrote the manuscript. Y.Y. and Y.Z. performed research, analyzed data, and drafted the manuscript; S.L., X.Z., M.W., and K.L. supervised the project. All authors read and approved the final manuscript.

Availability of data and materials

The datasets generated during and/or analysed during the current study are available in GEO database. Source code is available at https://github.com/bio-LIFTS/LIFTS.

REFERENCES

[1] D. B. Mlgaard-Nielsen, Pasternak, A. Hviid, Med, and Woensel, "Severe sepsis and septic shock," *New England Journal of Medicine*, vol. 369, no. 21, p. 2062, 2013.

[2] T. Poll, "Future of sepsis therapies," *Critical Care*, vol. 20, 12 2016.

[3] L. Weng, X.-y. Zeng, P. Yin, L.-j. Wang, C. Wang, W. Jiang, M.-g. Zhou, and B. Du, "Sepsis-related mortality in china: a descriptive analysis," *Intensive Care Medicine*, vol. 44, 05 2018.

[4] X. Zheng, K.-S. Leung, M.-H. Wong, and L. Cheng, "Long non-coding RNA pairs to assist in diagnosing sepsis," *BMC Genomics*, vol. 22, no. 1, apr 2021.

[5] B. Coburn, A. Morris, G. Tomlinson, and A. Detsky, "Does this adult patient with suspected bacteremia require blood cultures?" *JAMA : the journal of the American Medical Association*, vol. 308, pp. 502–11, 08 2012.

[6] S. Jain, D. Williams, S. Arnold, K. Ampofo, A. Bramley, C. Reed, C. Stockmann, E. Anderson, C. Grijalva, W. Self, Y. Zhu, A. Patel, W. Hymas, J. Chappell, R. Kaufman, J. Kan, D. Dansie, N. Lenny, D. Hillyard, and R. Rolfs, "Community-acquired pneumonia requiring hospitalization among u.s. children," *The New England journal of medicine*, vol. 372, pp. 835–45, 02 2015.

[7] A. Heffernan and K. Denny, "Host diagnostic biomarkers of infection in the icu: Where are we and where are we going?" *Current Infectious Disease Reports*, vol. 23, 04 2021.

[8] C. Pierrakos, D. Velissaris, M. Bisdorff, J. C. Marshall, and J.-L. Vincent, "Biomarkers of sepsis: time for a reappraisal," *Critical Care*, vol. 24, no. 1, jun 2020.

[9] C. Pierrakos, D. Velissaris, M. Bisdorff, J. Marshall, and J.-L. Vincent, "Biomarkers of sepsis: Time for a reappraisal," *Critical Care*, vol. 24, 12 2020.

[10] X. Liu, X. Zheng, J. Wang, N. Zhang, K. Leung, X. Ye, and L. Cheng, "A long non-coding rna signature for diagnostic prediction of sepsis upon icu admission," *Clinical and Translational Medicine*, vol. 10, 07 2020.

[11] L. Cheng, C. Nan, L. Kang, N. Zhang, S. Liu, H. Chen, C. Hong, Y. Chen, Z. Liang, and X. Liu, "Whole blood transcriptomic investigation identifies long non-coding rnas as regulators in sepsis," *Journal of Translational Medicine*, vol. 18, 05 2020.

[12] X. Liu, Y. Xu, R. Wang, S. Liu, J. Wang, Y. Luo, K. Leung, and L. Cheng, "A network-based algorithm for the identification of moonlighting noncoding rnas and its application in sepsis," *Briefings in bioinformatics*, 01 2020.

[13] C. chuan Nan, N. Zhang, K. C. P. Cheung, H. dong Zhang, W. Li, C. ying Hong, H. sheng Chen, X. yan Liu, N. Li, and L. Cheng, "Knockdown of lncRNA MALAT1 alleviates LPS-induced acute lung injury via inhibiting apoptosis through the miR-194-5p/FOXP2 axis," *Frontiers in Cell and Developmental Biology*, vol. 8, oct 2020.

[14] L. Cheng and Y. Luo, "An overview and metanalysis of machine and deep learning-based crispr grna design tools," *RNA biology*, 09 2019.

[15] J. Wang, X. Xiang, L. Bolund, X. Zhang, L. Cheng, and Y. Luo, "Gnl-scorer: A generalized model for predicting crispr on-target activity by machine learning and featurization," *Journal of molecular cell biology*, 01 2020.

[16] X. Zheng, Q. Wu, H. Wu, K.-S. Leung, M.-H. Wong, X. Liu, and L. Cheng, "Evaluating the consistency of gene methylation in liver cancer using bisulfite sequencing data," *Frontiers in Cell and Developmental Biology*, vol. 9, p. 1022, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fcell.2021.671302

[17] S. Liu, W. Zhao, X. Liu, and L. Cheng, "Metagenomic analysis of the gut microbiome in atherosclerosis patients identify cross-cohort microbial signatures and potential therapeutic target," *The FASEB Journal*, vol. 34, no. 11, pp. 14 166–14 181, sep 2020.

[18] J. Wang, X. Zhang, L. Cheng, and Y. Luo, "An overview and metanalysis of machine and deep learning-based CRISPR gRNA design tools," *RNA Biology*, vol. 17, no. 1, pp. 13–22, sep 2019.

[19] B. Scicluna, P. Klouwenberg, L. Vught, M. Wiewel, D. Ong, A. Zwinderman, M. Franitza, M. Toliat, P. Nürnberg, A. Hoogendijk, J. Horn, O. Cremer, M. Schultz, M. Bonten, and T. Poll, "A molecular biomarker to diagnose community-acquired pneumonia on intensive care unit admission," *American journal of respiratory and critical care medicine*, vol. 192, 06 2015.

[20] L. Mchugh, T. Seldon, R. Brandon, J. Kirk, A. Rapisarda, A. Sutherland, J. Presneill, D. Venter, J. Lipman, M. Thomas, P. Klouwenberg, L. Vught, B. Scicluna, M. Bonten, O. Cremer, M. Schultz, T. Poll, T. Yager, and R. Brandon, "A molecular host response assay to discriminate between sepsis and infection-negative systemic inflammation in critically ill patients: Discovery and validation in independent cohorts," *PLoS Medicine*, vol. 12, p. e1001916, 12 2015.

[21] B. Scicluna, M. Wiewel, L. Vught, A. Hoogendijk, A. Klarenbeek, M. Franitza, M. Toliat, P. Nürnberg, J. Horn, M. Bonten, M. Schultz, O. Cremer, and T. Poll, "A molecular biomarker to assist in diagnosing abdominal sepsis upon intensive care unit admission," *American Journal of Respiratory and Critical Care Medicine*, vol. 197, 10 2017.

[22] J. Wynn, C. Wilson, J. Hawiger, P. Scumpia, A. Marshall, J.-H. Liu, I. Zharkikh, H. Wong, P. Lahni, J. Benjamin, E. Plosa, J.-H. Weitkamp, E. Sherwood, L. Moldawer, R. Ungaro, H. Baker, M. C. Lopez, S. Mcelroy, N. Colliou, and D. Moore, "Targeting il-17a attenuates neonatal sepsis mortality induced by il-18," *Proceedings of the National Academy of Sciences*, vol. 113, p. 201515793, 04 2016.

[23] X. Liu, N. Li, S. Liu, J. Wang, N. Zhang, X. Zheng, K. Leung, and L. Cheng, "Normalization methods for the analysis of unbalanced transcriptome data: A review," *Frontiers in Bioengineering and Biotechnology*, vol. 7, 11 2019.

[24] L. Cheng, X. Wang, P. K. Wong, K.-Y. Lee, L. Li, B. Xu, D. Wang, and K. Leung, "Icn: A normalization method for gene expression data considering the over-expression of informative genes," *Mol. BioSyst.*, vol. 12, 10 2016.

[25] L. Cheng, L.-Y. Lo, N. Tang, D. Wang, and K. Leung, "Crossnorm: A novel normalization strategy for microarray data in cancers," *Scientific Reports*, vol. 6, p. 18898, 01 2016.

[26] Y. Li, Z. Lu, Y. Che, J. Wang, S. Shouguo, J. Huang, S. Mao, Y. Lei, Z. Chen, and J. he, "Immune signature profiling identified

predictive and prognostic factors for esophageal squamous cell carcinoma," *OncoImmunology*, vol. 6, 07 2017.

[27] P. Ghatalia, J. Gordetsky, F. Kuo, E. Dulaimi, K. Cai, K. Devarajan, S. Bae, G. Naik, T. Chan, R. Uzzo, A. Hakimi, G. Sonpavde, and E. Plimack, "Prognostic impact of immune gene expression signature and tumor infiltrating immune cells in localized clear cell renal cell carcinoma," *Journal for ImmunoTherapy of Cancer*, vol. 7, 12 2019.

[28] M. G. Mendel Suchmacher, *Practical Biostatistics*, 2nd ed. Academic Press, 2012, ch. 13, pp. 159–166.

[29] T. Li, R. Wernersson, R. Hansen, H. Horn, J. Mercer, G. Slodkowicz, C. Workman, O. Rigina, K. Rapacki, H. Stærfeldt, S. Brunak, T. Jensen, and K. Lage, "A scored human protein–protein interaction network to catalyze genomic interpretation," *Nature Methods*, vol. 14, 11 2016.

[30] L. Cheng and K. Leung, "Identification and characterization of moonlighting long non-coding rnas based on rna and protein interactome," *Bioinformatics*, vol. 34, 05 2018.

[31] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "clusterprofiler: an r package for comparing biological themes among gene clusters," *Omics : a journal of integrative biology*, vol. 16, pp. 284–7, 03 2012.

[32] C. Blatti, A. Emad, M. J. Berry, L. Gatzke, M. Epstein, D. Lanier, P. Rizal, J. Ge, X. Liao, O. Sobh, M. Lambert, C. S. Post, J. Xiao, P. Groves, A. T. Epstein, X. Chen, S. Srinivasan, E. Lehnert, K. R. Kalari, L. Wang, R. M. Weinshilboum, J. S. Song, C. V. Jongeneel, J. Han, U. Ravaioli, N. Sobh, C. B. Bushell, and S. Sinha, "Knowledge-guided analysis of "omics" data using the KnowEnG cloud platform," *PLOS Biology*, vol. 18, no. 1, p. e3000583, jan 2020.

[33] L. Cheng and K. Leung, "Quantification of non-coding rna target localization diversity and its application in cancers," *Journal of Molecular Cell Biology*, vol. 10, 01 2018.

[34] L. Cheng, P. Liu, and K. Leung, "Smile: A novel procedure for subcellular module identification with localization expansion," *IET Systems Biology*, vol. 12, 01 2018.

[35] L. Cheng, K. Fan, Y. Huang, D. Wang, and K. Leung, "Full characterization of localization diversity in human protein interactome," *Journal of Proteome Research*, vol. 16, 07 2017.

[36] L. Cheng, Y. Zeng, S. Hu, N. Zhang, K. C. P. Cheung, B. Li, K.-S. Leung, and L. Jiang, "Systematic prediction of autophagy-related proteins using arabidopsis thaliana interactome data," *The Plant Journal*, vol. 105, no. 3, pp. 708–720, dec 2020.

[37] L. Li, M. Liu, L. Yue, R. Wang, N. Zhang, Y. Liang, L. Zhang, L. Cheng, J. Xia, and R. Wang, "Host–guest protein assembly for affinity purification of methyllysine proteomes," *Analytical Chemistry*, vol. 92, no. 13, pp. 9322–9329, jun 2020.

[38] L. Cheng, P. Liu, D. Wang, and K. Leung, "Exploiting locational and topological overlap model to identify modules in protein interaction networks," *BMC Bioinformatics*, vol. 20, 01 2019.

[39] R. Yin, X. Liu, J. Yu, J. Yingbin, L. Jian, L. Cheng, and J. Zhou, "Up-regulation of autophagy by low concentration of salicylic acid delays methyl jasmonate-induced leaf senescence," *Scientific Reports*, vol. 10, p. 11472, 07 2020.

[40] X. Liu, X. Zheng, J. Wang, N. Zhang, K.-S. Leung, X. Ye, and L. Cheng, "A long non-coding RNA signature for diagnostic prediction of sepsis upon ICU admission," *Clinical and Translational Medicine*, vol. 10, no. 3, jul 2020.

[41] B. P. Scicluna, L. A. van Vught, A. H. Zwinderman, M. A. Wiewel, E. E. Davenport, K. L. Burnham, P. Nürnberg, M. J. Schultz, J. Horn, O. L. Cremer, M. J. Bonten, C. J. Hinds, H. R. Wong, J. C. Knight, T. van der Poll, F. M. de Beer, L. D. Bos, J. F. Frencken, M. E. Koster-Brouwer, K. van de Groep, D. M. Verboom, G. J. Glas, R. T. van Hooijdonk, A. J. Hoogendijk, M. A. Huson, P. M. K. Klouwenberg, D. S. Ong, L. R. Schouten, M. Straat, E. Witteveen, and L. Wieske, "Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study," *The Lancet Respiratory Medicine*, vol. 5, no. 10, pp. 816–826, oct 2017.

[42] R. Tibshirani, T. Hastie, and J. Friedman, "Regularized paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, 02 2010.

Yueran Yang received her BSc. in Mathematics and Applied Mathematics from Shanghai Jiao Tong University. She is currently pursuing her Master's degree at Cornell University. Her research interests include data mining, mathematical modelling and online learning.

Yu Zhang received his BSc. in Mathematics and Applied Mathematics from Shanghai Jiao Tong University. He is currently pursing his Master's degree at Nanyang Technological University. His research interests include data mining, 3D computer vision and natural language processing.

Shuai Li is currently a tenure-track assistant professor at John Hopcroft Center of Shanghai Jiao Tong University. She received Ph.D. in computer science and engineering from the Chinese University of Hong Kong. Before that, she received a bachelor's degree in Mathematics from Zhejiang University and a master's degree in Mathematics from the University of the Chinese Academy of Sciences. She has published many top conference papers on ICML/NeurIPS/AAAI/KDD/IJCAI/etc. and serves as reviewers on these conferences. She has visited/interned at many top universities and research labs like UC Berkeley/ University of Alberta/Microsoft/Huawei/Adobe/DeepMind/Tencent AI Lab/etc. She is one of the recipients of Google Ph.D. Fellowship in 2018.

Xubin Zheng received the BS degrees from Zhejiang University, China, in 2014 and MS degrees from Hong Kong University in 2016. Currently he is working toward the doctoral degree in the department of computer science and engineering, the Chinese University of Hong Kong. His research interests include bioinformatics, artificial intelligence, and data mining.

Man-Hon Wong received his BS and MP degrees from The Chinese University of Hong Kong in 1987 and 1989 respectively. He got the Ph.D. degree in University of California at Santa Barbara in 1993. Currently he is an associate professor at the department of computer science and engineering, The Chinese University of Hong Kong. His research interests include transaction management, mobile Databases, data replication, distributed systems, expert systems and applications of fuzzy logic.

Kwong-Sak Leung received his B.Sc. and Ph.D. degrees from the University of London in 1977 and 1980 respectively. He is currently Research Professor and appointed as Professor in the CUHK-BGI Innovation Institution of Trans-omics in the Chinese University of Hong Kong. His research interests include bioinformatics, artificial intelligence, and data mining.

Lixin Cheng received BS and MS degrees from Harbin Medical University, China, in 2008 and 2011, and PhD degree from Department of Computer Science & Engineering at the Chinese University of Hong Kong in 2018. Currently he is working as a PI at Shenzhen People's Hospital, First Affiliated Hospital of Southern University of Science and Technology, China. His research interests include bioinformatics, computational biology, and machine learning.

TABLE 1: **Summary of the gene expression cohorts used in this study.**

Series	Gene Number	Normal	Sepsis	Cell type	Age	Platform
Discovery Cohort I						
GSE57065	23521	25	82	Whole blood	Adult	
Discovery Cohort II						
GSE26378	23521	21	82	Whole blood	Children	
Validation Cohorts I						
GSE95233	23521	22	102	Whole blood	Adult	
GSE28750	23521	20	10	Whole blood	Adult	Affymetrix Human Genome U133 Plus 2.0
GSE8121	23521	15	60	Whole blood	Children	
GSE13904	23521	18	52	Whole blood	Children	
GSE26440	23521	32	98	Whole blood	Children	
GSE9692	23521	15	30	Whole blood	Children	
GSE4607	23521	15	69	Whole blood	Children	
Validation Cohorts II						
GSE65682	19040	42	479	Whole blood	Adult	Affymetrix Human Genome U219 Array
E-MTAB-1548	17028	15	80	Peripheral blood	Adult	Agilent Human Gene Expression 4x44K v2 Microarray

TABLE 2: **Genome characteristics of the genes in LIFTS.**

| Gene symbol	Gene name	Alignments	Chromosomal Location	Degree
LRRN3	leucine rich repeat neuronal 3	chr7:110731149-110765507 (+)	chr7q31.1	1
IL2RB	interleukin 2 receptor, beta	chr22:37521886-37545962 (-)	chr22q13.1	45
FCER1A	Fc fragment of IgE, high affinity I, receptor for; alpha polypeptide	chr1:159272125-159277991 (+)	chr1q23	19
TLR5	toll-like receptor 5	chr1:223283646-223316624 (-)	chr1q41-q42	7
S100A12	S100 calcium binding protein A12	chr1:153346183-153348075 (-)	chr1q21	3
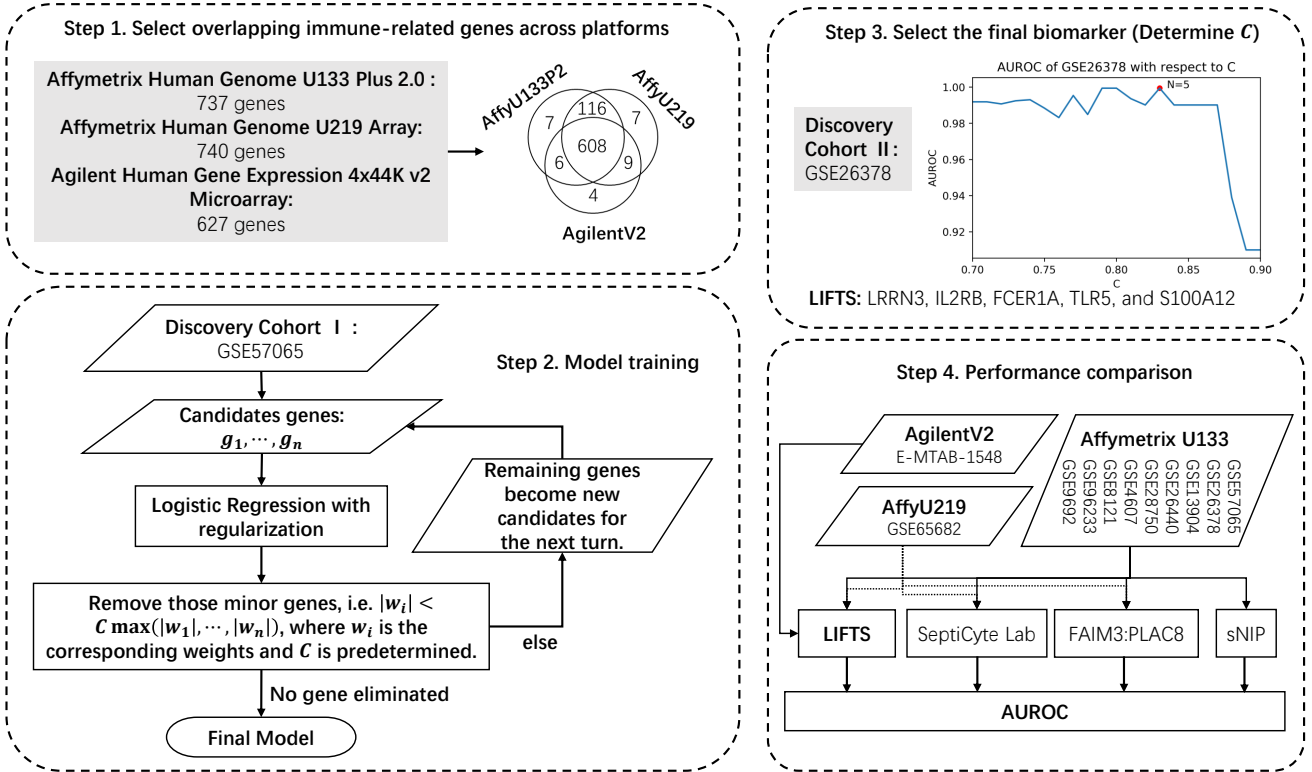
Figure 1: Workflow for the identification of sepsis biomarkers. a) statistics of immune-related genes of all cohorts in three platforms. b) the flow chart of the recurrent logistic regression algorithm including the regression step and the elimination step. c) determining the hyperparameter c with another discovery cohort and finalizing the biomarkers. d) validating and comparing diagnostic capability with distinct cohorts and platforms.
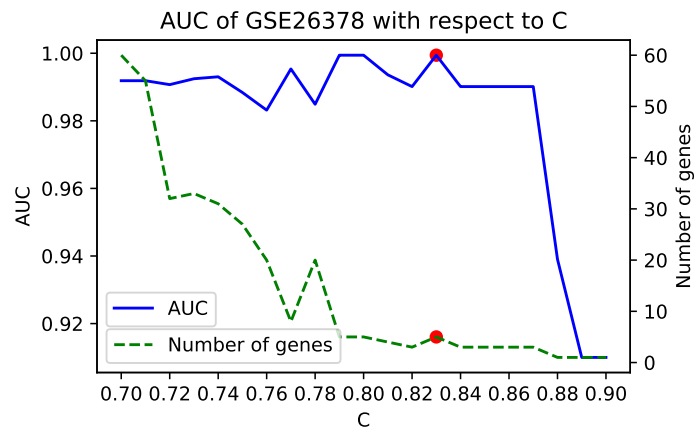
Figure 2: The solid blue line is the AUC with respect to different hyperparameter C between 0.7 and 0.99 with an interval of 0.01. The dash green line is the number of genes with respect to different hyperparameter C between 0.7 and 0.99 with an interval of 0.01. The red dot represents the optimal model with the corresponding $C = 0.83$, $gene\_num = 5$ and $AUC = 0.9994$.
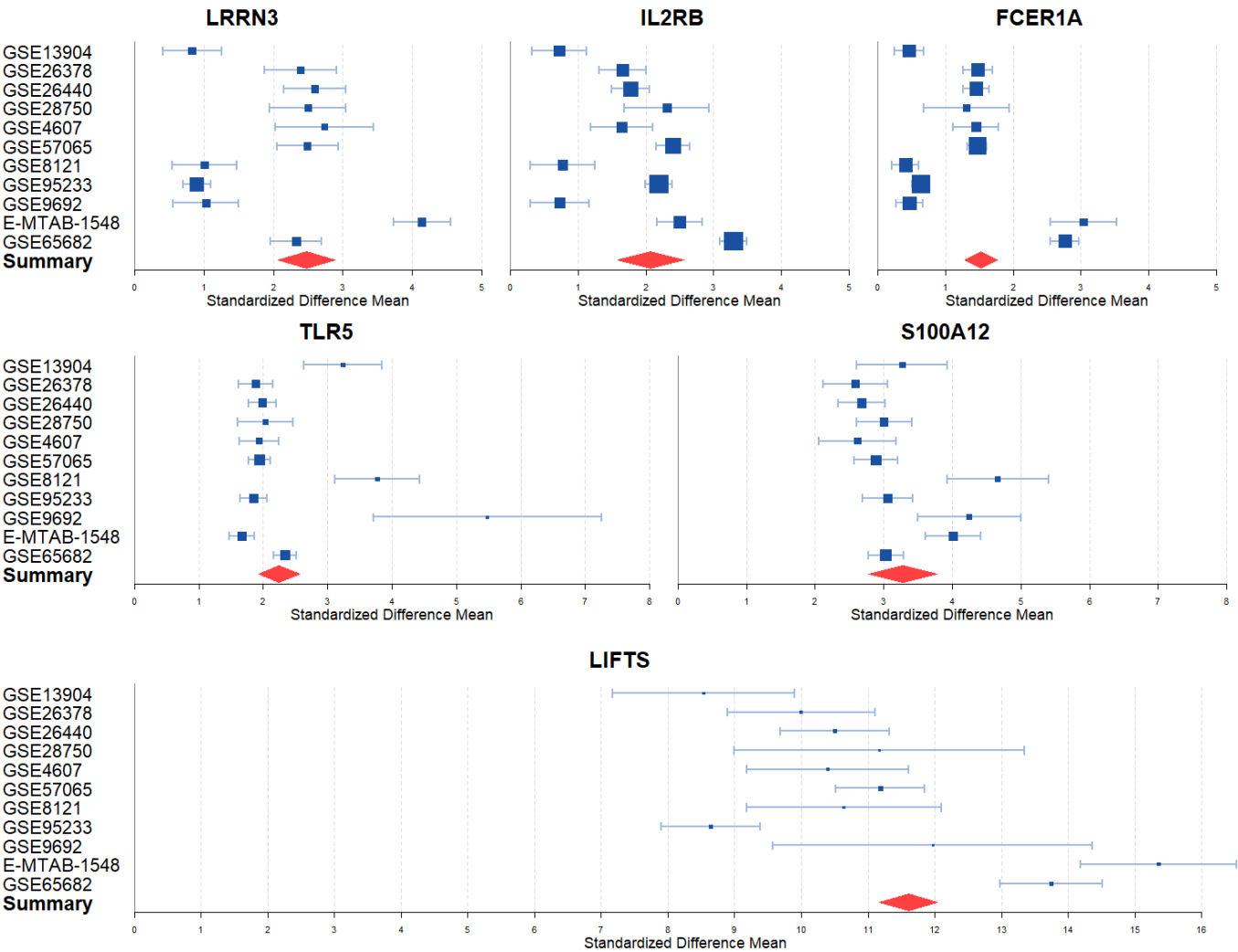
Figure 3: Forest plots of LIFTS and each genes in LIFTS. The x-axis represents the standardized mean difference between sepsis patients and healthy controls. The blue square is the average value of the difference and the size corresponds to the concentration of the data. The blue line represents the 95% confidence interval. The red diamond represents the average difference of a given gene or LIFTS for all cohorts. The width of the diamond represents the 95% confidence interval of the overall mean difference.
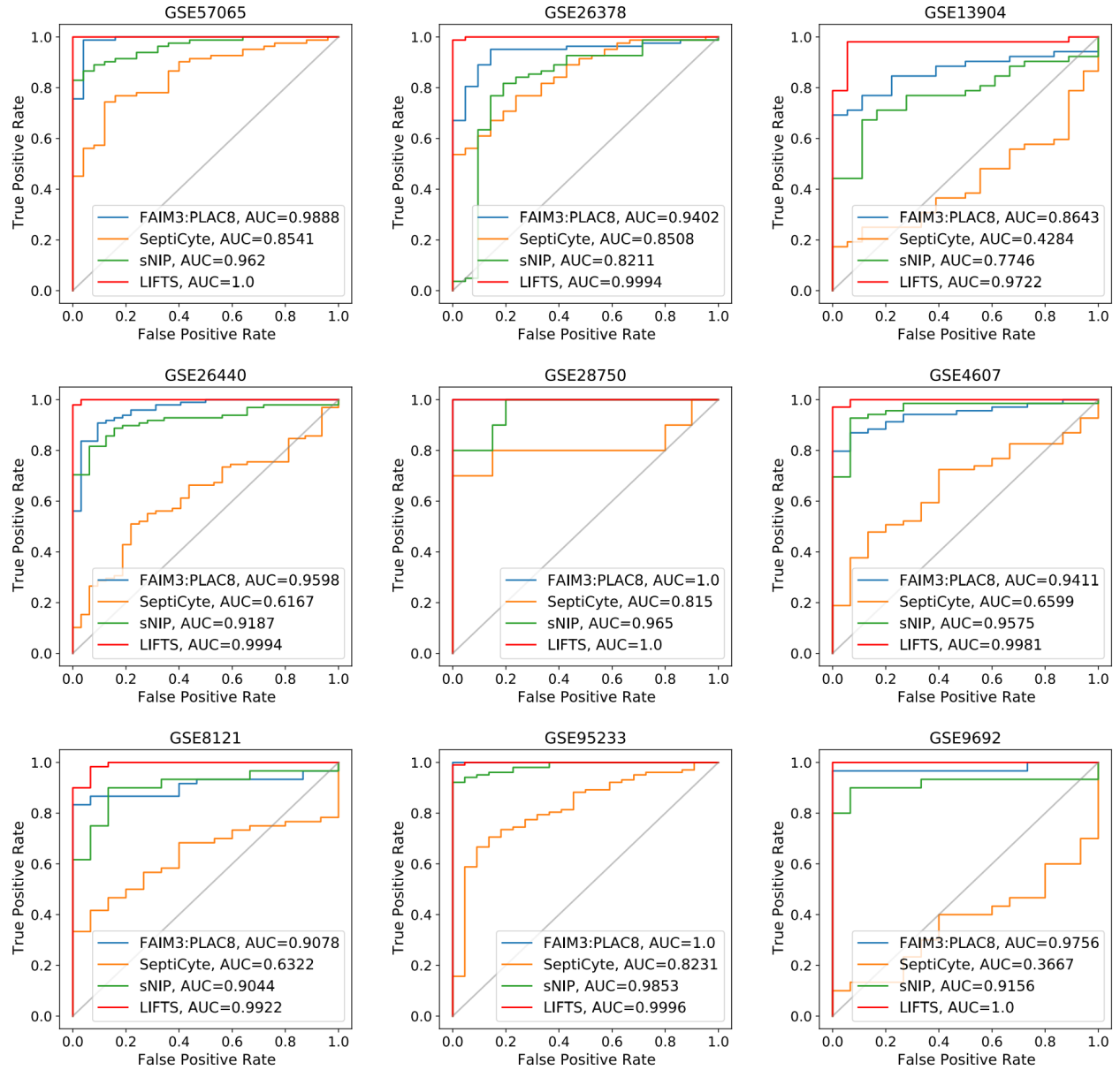
Figure 4: Performance comparison of LIFTS and other existing models in the discovery and validation cohorts. The first two cohorts, GSE57065 and GSE26378, are the discovery cohorts, while the others are validation cohorts.
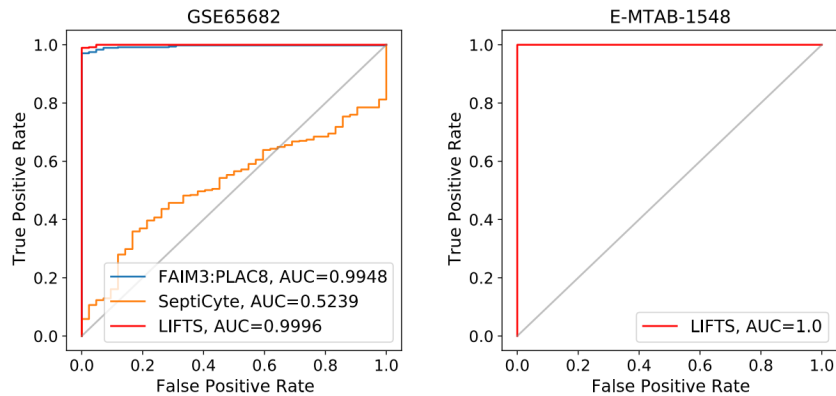

Figure 5: The performance of LIFTS based on two independent cohorts and platforms. NLRP1 does not exist on GSE65682 and PLAC8 is not available on E-MTAB-1548, resulting in the absence of some biomarkers.

Figure 6: Network and functional analysis of LIFTS. (A) Protein-protein interaction network of LIFTS and their interactors. (B) Interacting degree of LIFTS in the network. (C) Top 20 biological functions overrepresented by genes in the network. (D) Density distribution of simulated networks. (E) Function analysis using the pipeline of network-guided gene set characterization in KnowEnG.