A Robust and Generalizable Immune-Related Signature for Sepsis Diagnostics

Yueran Yang, Yu Zhang, Shuai Li, Xubin Zheng, Man-Hon Wong, Kwong-Sak Leung, and Lixin Cheng

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 utility 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 [8].

In recent years, with the rise of high-throughput sequencing technology, tens of thousands of genes can be detected in parallel [9], [10], [11], [12], providing opportunities for precise molecular diagnosis using machine learning methods [13], [14], [15], [16]. 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) [17], 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 [18]. Scicluna et al. developed a sNIP score for sepsis diagnosis based on the expression abundance of three genes using SVM and bootstrapping [19]. 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 [18], [19], [20], 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 [21], [22], [23]. 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.[24], [25] 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 (Fig. 1a).

2.3 Recurrent Logistic Regression

Recurrent logistic regression contains many iterations with model optimization and automatic feature selection, since each iteration involves regression step and elimination step (Fig. 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

\[ x = (g_1, \ldots, g_n)^T, \]  

where \( g_i \) is the \( i \)th gene expression.
To construct a classifier involving fewer genes features based on the expression vector $x$ 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

$$f(x) = \frac{1}{1 + e^{-w^T x}} = \frac{1}{1 + e^{-\sum_{i=0}^{n} w_i g_i}},$$

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

$$J(w; X) = \frac{1}{2} w^T w + \sum_{i=1}^{n} \log \left( 1 + \exp\left( -y_i (x_i^T w + w_0) \right) \right),$$

where $X$ is the collection of samples $x_1, x_2, \ldots, x_n$ 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.,

$$|w_i| < C \max(|w_1|, \ldots, |w_n|),$$

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 (Fig. 1c). We exhaustively 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 discriminative 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 (Fig. 3). Four graphical elements are presented including the estimated SMD (solid block), the
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 molecular events involved in sepsis. The recurrent logistic regression 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. Fig. 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

$$y(x) = [1 + \exp(-0.9305g_{LRRN3} - 0.9692g_{IL2RB} - 0.7378g_{IL2RB} + 0.8460g_{TLR5} + 0.8905g_{S100A12} - 0.0153)]^{-1},$$

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 corresponding part in logits of our diagnostic model to represent the diagnostic ability of each gene and LIFTS. Specifically, the logit is

$$\text{logit} = -0.9305g_{LRRN3} - 0.9692g_{IL2RB} - 0.7378g_{FCER1A} + 0.8460g_{TLR5} + 0.8905g_{S100A12}.$$  (6)

The standard difference mean $\bar{X} - \bar{Y}$ in effect size between the sepsis and control subjects is displayed in Fig. 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. Fig. 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

| 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 (+) | chr1q25            | 19     |
| TLR5        | toll-like receptor 5        | chr1:223283646-223316624 (-) | chr1q41-q42          | 7      |
| S100A12     | S100 calcium binding protein A12 | chr1:153346183-153348075 (-) | chr1q21            | 3      |
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 is 0.9994 and 1.0, which are superior to its counterparts, indicating the portability of LIFTS among different platforms in diagnostic prediction (Fig. 5).

The AUC curves are related to the standard difference means. Considering LIFTS performance shown in Figs. 3 and 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 [31], [32], [33], [34]. The interference in protein interactions and new undesired protein interactions can cause diseases [35], [36], [37]. 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 (Fig. 6 A). Specifically, 1, 45, 19, 7, and 3 partner genes interact with LRRN3, IL2RB, FCER1A, TLR5, and S100A12, respectively (Fig. 6 B). 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. (Fig. 6 C). For the protein interaction network, the connections among genes are significantly dense than the simulated networks ($P < e^{-26}$, Rank sum test, Fig. 6 D), where we randomly picked up the same number of proteins

Fig. 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.
Fig. 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.

Fig. 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.
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 [30] 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 megakaryocytes 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 [19]. 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 [9], [38]. 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 Figs. S2 and S3, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2021.3107874), 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 Fig. S1, available online).

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 [39]. 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.

ACKNOWLEDGMENTS

This work was supported by the Basic and Applied Basic Research Programs Foundation of Guangdong Province under Grant 2019A1515110097. 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. Yueran Yang and Yu Zhang are contribute equally to this study.

REFERENCES

[1] J. Calis, J. van Woensel, and J. Lemson, “Severe sepsis and septic shock,” New Engl. J. Med., vol. 369, no. 21, 2013, Art. no. 2062.
[2] T. Poll, “Future of sepsis therapies,” Crit. Care, vol. 20, 2016, Art. no. 106.
[3] L. Weng et al., “Sepsis-related mortality in China: A descriptive analysis,” Intensive Care Med., vol. 44, no. 7, pp. 1071–1080, 2018.
[4] X. Zheng, K.-S. Leung, M.-H. Wong, and L. Cheng, “Long non-coding RNA pairs to assist in diagnosing sepsis,” BMC Genomic., vol. 369, no. 21, 2013, Art. no. 2062.
[5] B. Coburn, A. Morris, G. Tomlinson, and A. Detsky, “Does this adult inpatient with suspected bacteremia require blood cultures?”, JAMA - J. Amer. Med. Assoc., vol. 308, no. 5, pp. 502–511, 2012.
[6] S. Jain et al., “Community-acquired pneumonia requiring hospitalization among U.S. children,” New Engl. J. Med., vol. 372, pp. 835–845, 2015.
[7] A. Heffernan and K. Denny, “Host diagnostic biomarkers of infection in the ICU: Where are we and where are we going?”, Curr. Infect. Dis. Rep., vol. 23, 2021, Art. no. 4.
[8] C. Pierrakos, D. Velissaris, M. Bisdorff, J. C. Marshall, and J.-L. Vincent, “Biomarkers of sepsis: Time for a reappraisal,” Crit. Care, vol. 24, no. 1, 2020, Art. no. 287.
[9] X. Liu et al., “A long non-coding RNA signature for diagnostic prediction of sepsis upon ICU admission,” Clin. Transl. Med., vol. 10, no. 3, 2020, Art. no. e123.
[10] L. Cheng et al., “Whole blood transcriptomic investigation identifies long non-coding RNAs as regulators in sepsis,” J. Transl. Med., vol. 18, 2020, Art. no. 217.
[11] X. Liu et al., “A network-based algorithm for the identification of moonlighting noncoding RNAs and its application in sepsis,” Brief. Bioinf., vol. 22, no. 1, pp. 581–588, 2020.
[12] C. C. Nan et al., “Knockdown of lncRNA MALAT1 alleviates LPS-induced acute lung injury via inhibiting apoptosis through the miR-194-5p/FOPX2 axis,” Front. Cell. Dev. Biol., vol. 8, 2020, Art. no. 586869.
[13] L. Cheng and Y. Luo, “An overview and metanalysis of machine and deep learning-based CRISPR gRNA design tools,” RNA Biol., vol. 17, no. 1, pp. 13–22, 2020.
[14] J. Wang, X. Xiang, L. Bolund, X. Zhang, L. Cheng, and Y. Luo, “GNI-scorer: A generalized model for predicting CRISPR on-target activity by machine learning and featureization,” J. Mol. Cell Biol., vol. 12, no. 11, pp. 909–911, 2020.
[15] X. Zheng et al., “Evaluating the consistency of gene methylation in liver cancer using bisulite sequencing data,” Front. Cell. Dev. Biol., vol. 9, 2021, Art. no. 671302.
[16] S. Liu, W. Zhao, X. Liu, and L. Cheng, “Metagenomic analysis of the gut microbiome in atherosclerosis patients identify cross-collaboration microbial metabolic and potential therapeutic target,” FASEB J., vol. 34, no. 11, pp. 14166–14181, 2020.
[17] B. Scicluna et al., “A molecular biomarker to diagnose community-acquired pneumonia on intensive care unit admission,” Amer. J. Respir. Crit. Care Med., vol. 192, no. 7, pp. 827–835, 2015.
[18] L. Mchugh et al., “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 Med., vol. 12, 2015, Art. no. e1001916.
[19] B. Scicluna et al., “A molecular biomarker to assist in diagnosing abdominal sepsis upon intensive care unit admission,” Amer. J. Respir. Crit. Care Med., vol. 197, no. 8, pp. 1070–1073, 2017.
[20] J. Wynn et al., “Targeting IL-17A attenuates neonatal sepsis mortality induced by IL-18,” Proc. Nat. Acad. Sci. USA, vol. 113, no. 19, pp. 5267–5265, 2016.
[21] X. Liu et al., “Normalization methods for the analysis of unbalanced transcriptome data: A review,” Front. Bioeng. Biotechnol., vol. 7, 2019, Art. no. 358.
Yueran Yang received the BSc degree in mathematics and applied mathematics from Shanghai Jiao Tong University. She is currently working toward the master’s degree with Cornell University. Her research interests include data mining, mathematical modelling, and online learning.

Yu Zhang received the BSc degree in mathematics and applied mathematics from Shanghai Jiao Tong University. He is currently working toward the master’s degree with Nanyang Technological University. His research interests include data mining, 3D computer vision, and natural language processing.

Shuai Li received the bachelor’s degree in mathematics from Zhejiang University, the master’s degree in mathematics from the University of the Chinese Academy of Sciences and the PhD degree in computer science and engineering from The Chinese University of Hong Kong. She is currently a tenure-track assistant professor with John Hopcroft Center, Shanghai Jiao Tong University. She has visited or interned at many top universities and research labs, including UC Berkeley, University of Alberta, Microsoft, Huawei, Adobe, DeepMind, and Tencent AI Lab. She has authored and coauthored many top conference papers, including the ICML, NeurIPS, AAAI, KDD, and IJCAI, she is also the reviewer of these conferences. She was one of the recipients of the Google PhD Fellowship in 2018.

Xubin Zheng received the BS degree from Zhejiang University, China, in 2014 and the MS degree from Hong Kong University in 2016. He is currently working toward the doctoral degree with 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 the BS and MP degrees from The Chinese University of Hong Kong, in 1987 and 1989, respectively, and the PhD degree from the University of California, Santa Barbara in 1993. He is currently an associate professor with the Department of Computer Science and Engineering, Chinese University of Hong Kong. His research interests include data management, mobile databases, data replication, distributed systems, expert systems, and applications of fuzzy logic.

Kwong-Sak Leung received the BSc and PhD degrees from the University of London, in 1977 and 1980, respectively. He is currently a research professor and a professor with the CUHK-BGI Innovation Institution of Trans-omics, Chinese University of Hong Kong. His research interests include bioinformatics, artificial intelligence, and data mining.

Lixin Cheng received the BS and MS degrees from Harbin Medical University, China, in 2008 and 2011, respectively, and the PhD degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, in 2018. He is currently a PI with 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.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.