Development and initial validation of diagnostic gene signatures for systemic lupus erythematosus

Systemic lupus erythematosus (SLE) is a complex and heterogeneous rheumatic disease with variable clinical features. The correct diagnosis of SLE is still challenging, partially due to the complexity and heterogeneity of SLE pathogenesis. A new classification criteria for SLE with excellent sensitivity and specificity has been recently proposed by the European League Against Rheumatism and the American College of Rheumatology.¹ Apart from autoantibodies such as anti-double-stranded DNA antibodies, novel molecular biomarkers may help to improve the performance of SLE classification criteria, but they are not included in the new classification criteria, which is largely attributed to the limited availability in the clinical setting or insufficient evidence.¹ Transcriptome studies using either microarray or RNA sequencing mainly aim to investigate the aberrant RNA expression levels of genes on a genome-wide scale, and have been widely used in the clinical research of rheumatic diseases.

Studies using transcriptome analysis also have provided deeper insights into the pathogenic mechanism of SLE. Differently expressed genes (DEGs) derived from microarray-based peripheral blood transcriptome data can also be used as diagnostic or prognostic biomarkers for autoimmune diseases, but their use in SLE is still not well established.^{2,3} In this study, we aimed to develop a genetic signature for SLE diagnosis through bioinformatic analyses of whole blood transcriptome data.

To overcome the limited clinical utility caused by the low consistency and the risk of noise discovery in microarray-based studies, robust rank aggregation (RRA) analysis was used to integrate data from multiple transcriptome datasets. RRA is a useful integration approach in pooling data from heterogeneous datasets and can help to identify the mostly aberrantly expressed genes between patients with SLE and controls across multiple datasets, which may lead to a SLE diagnostic gene signature with both high reproducibility and high stability.⁴ Fourteen whole blood transcriptome datasets with at least 20 patients with SLE and 10 controls were integrated with RRA, which included GSE110685, GSE112087, GSE99967, GSE110169, GSE88884, GSE65391, GSE72509, GSE45291, GSE49454, GSE61635,

Figure 1 Assessment of the diagnostic performance of gene signatures in systemic lupus erythematosus (SLE) in six microarray datasets. (A) Diagnostic performance of RRAtop10, RRAWGCNA10 and IFI44L alone in GSE45291 (292 patients with SLE and 20 controls). (B) Diagnostic performance of RRAtop10, RRAWGCNA10 and IFI44L alone in GSE49454 (157 patients with SLE and 20 controls). (C) Diagnostic performance of RRAtop10, RRAWGCNA10 and IFI44L alone in GSE61635 (79 patients with SLE and 30 controls). (D) Diagnostic performance of RRAtop10, RRAWGCNA10 and IFI44L alone in GSE65391 (118 patients with SLE and 32 controls). (E) Diagnostic performance of RRAtop10, RRAWGCNA10 and IFI44L alone in GSE88884 (1760 patients with SLE and 60 controls). (F) Diagnostic performance of RRAtop10, RRAWGCNA10 and IFI44L alone in GSE110169 (82 patients with SLE and 77 controls). Receiver operating characteristic curve analyses were performed using the enrichment score of RRAtop10 and RRAWGCNA10 in each individual calculated through Gene Set Variation Analysis. AUC, area under the curve.

Correspondence

Figure 2 Assessment of the diagnostic performance of gene signatures in systemic lupus erythematosus (SLE) in two datasets using RNA sequencing. (A) Diagnostic performance of RRAtop10, RRAWGCNA10 and IFI44L alone in GSE72509 (99 patients with SLE and 18 controls). (B) Diagnostic performance of RRAtop10, RRAWGCNA10 and IFI44L alone in GSE112087 (31 patients with SLE and 29 controls). (C) Comparison of SLE5genescore with single gene in diagnosing SLE in the discovery dataset GSE72509 (99 patients with SLE and 18 controls). (D) Comparison of SLE5genescore with single gene in diagnosing SLE in the validation dataset GSE112087 (31 patients with SLE and 29 controls). Receiver operating characteristic curve analyses were performed using the enrichment score of RRAtop10 and RRAWGCNA10 in each individual calculated through Gene Set Variation Analysis, and SLE5genescore was calculated through the log₂-transformed TPM expression values of five genes including IFI44L, CASP5, ANXA3, TCN1 and CXCR6. AUC, area under the curve.

GSE50635, GSE39088, GSE20864 and GSE17755. RRA outcomes suggested that most of those top 100 DEGs were from type I interferon-related pathways such as IFI44L, IFI27 and IFIT1. The top 10 upregulated genes included IFI44L, IFI27, RSAD2, IFIT1, HERC5, IFIT3, IFI44, OASL, CMPK2 and USP18 and were all from type I interferon-related pathways, which were preliminarily selected as one SLE diagnostic genetic signature referred to as RRAtop10 in this study.

Because gene signatures developed by combining genetic biomarkers from multiple functional modules may have the ability of leading to a more accurate diagnosis than gene signature derived from one single co-expression module,³ the co-expression pattern of those top 100 upregulated DEGs in GSE88884 (1760 patients with SLE and 60 controls) was further analysed using weighted gene coexpression network analysis (WGCNA).⁵ Based on the gene co-expression modules calculated above, a more complex gene signature was developed by selecting at most 10 genes from each upregulated co-expression module. This complex gene signature (referred to as RRAWGCNA10) consisted of 34 key genes from six independent co-expression modules and included IFI44L, IFI27, RSAD2, IFIT1, HERC5, IFIT3, IFI44, OASL, CMPK2, USP18, LHFPL2, RRM2, CEACAM6, CEACAM8, DEFA4, HP, LCN2, MMP8, OLFM4, OLR1, RNASE2, TCN1, ANKRD22, CASP5, CEACAM1, CLEC4D, DHRS9, DYNLT1, FCGR1B, TLR5, TNFAIP6, TNFSF13B, ANXA3 and SLC26A8. Among those 34 genes, the first 10 genes were identical to those genes in RRAtop10 and were all from the same co-expression module. Based on the

enrichment score of these genetic signatures in each individual calculated through Gene Set Variation Analysis, the diagnostic role of genetic signatures was assessed by receiver operating characteristic curve analysis.⁶ To evaluate the performance of those diagnostic gene signatures, the area under the curve (AUC) was calculated in six validation datasets using microarray (GSE45291, GSE49454, GSE61635, GSE65391, GSE88884 and GSE110169) and two validation datasets using RNA sequencing (GSE72509 and GSE112087). The performance of those two diagnostic gene signatures was compared with IFI44L, which was the top upregulated gene in the RRA analysis.

As shown in figures 1 and 2, as the most aberrantly expressed gene in the whole blood of patients with SLE, IFI44L alone could provide some assistance in diagnosing SLE with AUCs arranging from 0.79 and 0.94 among those six microarray datasets (figure 1A–F) and two RNA-sequencing datasets (figure 2A,B). Compared with IFI44L alone, RRAtop10 had a better performance in diagnosing SLE in only two datasets including GSE88884 ($p=0.02$; figure 1E) and GSE72509 ($p=0.02$; figure 2A), but not in the other six datasets such as GSE45291 ($p=0.35$; figure 1A) and GSE112087 ($p=0.10$; figure 2B). Compared with IFI44L alone, RRAWGCNA10 had a better performance in diagnosing SLE in four datasets including GSE49454 ($p=0.01$; figure 1B), GSE61635 ($p=0.03$; figure 1C), GSE88884 ($p=0.0001$; figure 1E) and GSE72509 ($p=0.03$; figure 2A), but not in the other four datasets such as GSE45291 ($p=0.80$; figure 1A) and GSE112087 ($p=0.18$; figure 2b). In addition, RRAWGCNA10 also had better

performance than RRAtop10 in diagnosing SLE in two datasets including GSE61635 ($p=0.04$; figure 1C) and GSE88884 ($p=0.005$; figure 1E), and had comparable performance with RRAtop10 in the other datasets. The outcomes above supported that gene signatures derived from whole blood transcriptional profiles could provide some assistance to the diagnosis of SLE, and gene signatures consisting of genes from multiple co-expression modules may have better diagnostic performance.

To improve the clinical application potential of SLE gene signatures, a simpler gene signature was developed using multivariate logistic regression analysis. Those genes analysed in the logistic regression analysis were key genes from main co-expression modules related to SLE in the WGCNA above, and only one candidate gene was selected from each module. Among those two RNA-sequencing datasets, GSE72509 (99 patients with SLE and 19 controls) was selected as the discovery dataset and GSE112087 (31 patients with SLE and 29 controls) was selected as the validation dataset. Owing to the obvious difference in data form between RNA-sequencing data and microarray data, datasets using microarray were not analysed in this part. Through logistic regression analysis in GSE72509, a five-gene diagnostic score for SLE (SLE5genescore) was developed through the log2-transformed transcripts per million (TPM) expression values of five genes including IFI44L, CASP5, ANXA3, TCN1 and CXCR6. The five-gene diagnostic score was calculated using the following formula: $0.76 \times \text{IFI44L} + 0.18 \times \text{CASP5} + 0.87 \times \text{ANXA3} + 0.74 \times \text{TCN1} - 0.53 \times \text{CXCR6}$. As shown in figure 2, SLE5genescore had an obvious better performance than single gene in diagnosing SLE in both the discovery dataset (GSE72509, AUC=0.95 (95% CI 0.90 to 0.99), $p<0.05$; figure 2C) and the validation dataset (GSE112087, AUC=0.93 (95% CI 0.87 to 0.99), $p<0.05$; figure 2D).

In summary, this study developed useful gene signatures for SLE diagnosis through bioinformatic analyses of whole blood transcriptomic data, which could effectively differentiate SLE in different datasets and may provide some assistance in the classification of SLE. Nevertheless, more studies are needed to further validate their performance in diagnosing SLE in different clinical settings.

Bin Wang , Shiju Chen, Qing Zheng, Zhenyu Gao, Rongjuan Chen, Jingxiu Xuan, Yuan Liu , Guixiu Shi

Correspondence to Dr Yuan Liu and Professor Guixiu Shi, Department of Rheumatology and Clinical Immunology, The First Affiliated Hospital of Xiamen University, Xiamen, China; liuyuan@xmu.edu.cn, gshi@xmu.edu.cn

Acknowledgements All authors would like to thank those researchers who had shared data in GEO database.

Contributors GS and BW performed the study design. BW and SC analysed data and wrote the manuscript. QZ, ZG, RC, JX and YL collected and analysed data. All authors approved the final manuscript.

Funding This work was supported by grants from the National Natural Science Foundation of China (Grant No. 81971536 and No. U1605223).

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; internally peer reviewed.

© Author(s) (or their employer(s)) 2019. No commercial re-use. See rights and permissions. Published by BMJ.

BW, SC and QZ contributed equally.

To cite Wang B, Chen S, Zheng Q, et al. *Ann Rheum Dis* Epub ahead of print: [please include Day Month Year]. doi:10.1136/annrheumdis-2019-216695

Received 22 November 2019

Accepted 27 November 2019

Ann Rheum Dis 2019;0:1–3. doi:10.1136/annrheumdis-2019-216695

ORCID iDs

Bin Wang <http://orcid.org/0000-0002-6985-8289>

Yuan Liu <http://orcid.org/0000-0003-2668-0350>

REFERENCES

- 1 Aringer M, Costenbader K, Daikh D, et al. European League against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. *Ann Rheum Dis* 2019;2019:1151–9.
- 2 Fricker M, Gibson PG, Powell H, et al. A sputum 6-gene signature predicts future exacerbations of poorly controlled asthma. *J Allergy Clin Immunol* 2019;144:51–60.
- 3 Giacomelli R, Afeltra A, Alunno A, et al. Guidelines for biomarkers in autoimmune rheumatic diseases—evidence based analysis. *Autoimmun Rev* 2019;18:93–106.
- 4 Kolde R, Laur S, Adler P, et al. Robust RANK aggregation for gene list integration and meta-analysis. *Bioinformatics* 2012;28:573–80.
- 5 Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;9:559.
- 6 Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* 2013;14:7.