Mol Cell Proteomics Papers in Press. Published on July 15, 2019 as Manuscript RA119.001582

Identification of serum biomarkers for Systemic Lupus Erythematosus using a library of phage displayed random peptides and deep sequencing

Fan-Lin Wu[1,3#], Dan-Yun Lai[1#], Hui-Hua Ding[4], Yuan-Jia Tang[4], Zhao-Wei Xu[1], Ming-Liang Ma[1], Shu-Juan Guo[1], Jingfang Wang[1], Nan Shen[4,5], Xiao-Dong Zhao[1] , Huan Qi[1*], Hua Li[2*]and Sheng-Ce Tao[1,2*]

[1]Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China

[2]Bio-ID Center, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China

[3]School of Agriculture, Ludong University, Yantai, 264025, China

[4]Shanghai Institute of Rheumatology, Department of rheumatology, Renji Hospital, Shanghai Jiao Tong University School of Medicine,145 Shan Dong Road (c), 200240, Shanghai, China

[5]State Key Laboratory of Oncogenes and Related Genes, Shanghai Cancer Institute, Renji Hospital, 2200 Lane 25 Xietu Road, 200240, Shanghai, China

[#]These authors contributed equally to this work

*Corresponding author. E-mail: taosc@sjtu.edu.cn or kaikaixinxin@sjtu.edu.cn or qihuan@sjtu.edu.cn

Abstract

Systemic lupus erythematosus (SLE) is one of the most serious autoimmune diseases, characterized by highly diverse clinical manifestations. Biomarker is still needed for the accurate diagnostics. SLE serum autoantibodies were discovered and validated using serum samples from independent sample cohorts encompassing 306 participants divided into three groups, *i.e.*, healthy, SLE patients, and other autoimmune related diseases. To discover biomarkers for SLE, a phage displayed random peptide library (Ph.D. 12) and deep sequencing were applied to screen specific autoantibodies in a total of 100 serum samples from 50 SLE patients and 50 healthy controls. A statistical analysis protocol was setup for the identification of peptides as potential biomarkers. For validation, ten peptides were analyzed using enzyme linked immunosorbent assays (ELISA). As a result, four peptides (SLE2018Val001, SLE2018Val002, SLE2018Val006 and SLE2018Val008) were discovered with high diagnostic power to differentiate SLE patients from health controls. Among them, two peptides, *i.e.*, SLE2018Val001 and SLE2018Val002 were confirmed between SLE with other autoimmune patients. The procedure we established could be easily adopted for the identification of autoantibodies as biomarkers for many other diseases.

Keywords: Phage display, Random peptide library, Systemic Lupus Erythematosus, Serum biomarker

Abbreviations: Systemic lupus erythematosus (SLE), Enzyme linked immunosorbent assays (ELISA), Anti-nuclear antibodies (ANA), Rheumatoid arthritis (RA), Dermatomyositis (DM), Behçet's disease (BD), Ankylosing spondylitis (AS)

Introduction

Systemic lupus erythematosus (SLE) is a chronic, complex autoimmune disorder characterized by the production of autoantibodies and heterogeneous clinical presentation (1). The incidence rate of SLE is 2-7/10,000 (2). It primarily affects women of childbearing age with a female-to-male ratio of 9:1 (3).

The clinical manifestations of SLE are diverse, including arthritis, rashes, nephritis, and serositis, which lead to reduced physical function, high morbidity, impaired quality of life, and shortened life span (4). Furthermore, both serological and immunological indicators of the disease are highly variable. Owing to its complexity and heterogeneity regarding its etiology, pathogenesis, and clinical presentation, SLE remains one of the greatest challenges for physicians to diagnose.

Loss of immune tolerance leading to immune dysregulation and excess production of autoantibodies causes the clinical manifestation of SLE. Autoantibodies are the hallmarks of SLE. More than 180 autoantibodies have been reported to be related to lupus (5). Autoantibodies have long been applied as biomarkers for the diagnosis of SLE. The most commonly used biomarker is anti-nuclear antibody (ANA), which has high sensitivity for the diagnosis of SLE, but it is not specific since it can be observed in a variety of other autoimmune diseases, such as rheumatoid arthritis (RA), systemic sclerosis, and autoimmune hemolytic anemia (6). Other autoantibodies such as anti-double stranded DNA or anti-Smith (Sm) are more specific for SLE but exist in only a fraction of lupus patients (7). Therefore, there is an urgent need for novel highly sensitive and specific SLE biomarkers that can be applied for the

diagnosis of SLE. A good SLE biomarker should be able to accurately differentiate SLE from other autoimmune diseases, and the diagnosis process should also be reliable, cost-effective and have no adverse effects on the patients. In 2013, using a library of synthetic autoantigen surrogates, Quan *et al.* identified an SLE marker with a specificity of 97.5% a sensitivity of 70% (8). In addition to chemically synthesized peptoids (9), proteins (10) and polypeptides (11) are also available as candidates for biomarkers.

Several approaches have been attempted to discover specific antibodies associated with autoimmune diseases. Because the content of protein in serum is heterogeneous and the abundance of autoantibodies tends to be low, it is difficult to solve the problem of significant differences in autoantibody levels in different samples by mass spectrometry. To overcome the drawbacks of conventional mass spectrometry methods, Hecker M *et al.* analyzed IgG autoantibody reactivity in serum and cerebrospinal fluid samples from multiple sclerosis patients with a high-density peptide microarray (12). Zhu *et al*. used an autoantigen microarray for high-throughput autoantibody profiling in SLE (13). However, due to the high cost of the array and library, the microarray based strategy for serum biomarker discovery has not been widely applied. Furthermore, most of the approaches have focused on a panel of antigens known or predicted to be important for the disease (14-16). Alternatively, phage display peptide libraries could also serve as only a handful potential source of antigens for autoantibody based serum biomarker discovery. Larman *et al*. constructed a T7-peptipe library containing the complete human proteome for autoantigen discovery. The binding peptides were then decoded by deep sequencing, and this technology was named PhIP-seq (17).

There are many diseases, including autoimmune diseases, for which the antigens triggering the primary immune response are unknown; thus, a definitive antigen is not available (9). In this scenario, a highly diversified library of random peptides may serve as a source of potential antigens. For example, Christiansen *et al*. used a random peptide library of 7 amino acids coupled with deep sequencing to identify possible IgE epitopes and allergens in peanut allergy patients. However, several rounds of screening were still required, given that only a handful of samples were tested, and no further validation was presented (18).

Herein, we demonstrate that an M13 phage displayed random peptide library (19) in combination with deep sequencing can be used to identify peptides that could be specifically recognized by IgG in the sera of SLE patients. For the screening phase, compared to that of the healthy controls, a total of 116 peptides were found to be highly enriched in the sera of SLE patients. Further validation showed that by using a four-peptide panel, an AUC of 0.86 could be achieved. Our methodology provides a novel unbiased approach for the discovery of serum biomarkers with low cost and high efficiency.

Experimental Procedures

Study Population and Serum sample preparation

The study consisted of a screening phase and two validation phases. In the screening phase, 50 SLE patients and 50 age and gender matched healthy controls were included. In validation phase I, 60 SLE patients and 60 age and gender matched healthy controls were included. In validation phase II, 15 SLE patients, 7 Rheumatoid arthritis (RA) patients, 9 Dermatomyositis

(DM) patients, 20 Behçet's disease (BD) patients, and 15 Ankylosing spondylitis (AS) patients were included. All the SLE patients fulfilled the 1997 American College of Rheumatology (ACR) classification criteria for SLE (20). The demographic characteristics of the patients and healthy controls were listed in **Tables I**, **II**, and **III**. Informed consent were obtained from all the participants and the study was approved by ethics committee of Shanghai Jiao Tong University school of medicine, Renji Hospital. All the sera samples were collected from Shanghai Renji Hospital and Shanghai Ruijin Hospital following the same standard operation procedure. The sera were prepared according to standard protocol (21).

Phage immuno-precipitation (IP)

The phage displayed peptide library (Ph.D. 12) was purchased from NEB and stored at -80°C until used. The concentration of IgG from the samples were measured with the commercial Human IgG ELISA Quantitation kits (eBioscience, MA, USA) according to the instructions. For all samples, the final amount of IgG added to 100 μL IP mix was 0.4 μg. Besides the serum samples, we also added 1 μg of an anti-6×His antibody (Millipore, CA, USA) and 10 μL PBS buffer to a 100 μL IP mix as positive control and negative control, respectively. About 1×10^9 M13-phage particles were added to 100 μL IP mix in a 96 well PCR plate. The plate was then carefully sealed with adhesive optical tape (Applied Biosystems, Carlsbad, USA) and rotating overnight at 4°C. 10 μL of protein G Dynabeads (Invitrogen, MA, USA) was then added to each well. The re-sealed plate was placed back on a rotator for 4 h at 4°C. The beads were washed 6 times in 200 μL wash buffer (150 mM NaCl, 50 mM Tris-HCl, 0.1% NP-40 (pH 7.5))

by pipetting up and down eight times per wash. After removal of the wash buffer, the beads were resuspended in 30 μL ddH_2O and heated at 95°C for 10 min.

Library preparation for deep sequencing

We prepared the library for deep sequencing with two rounds of PCR amplification of the phage DNA using hot startQ5 polymerase (NEB, MA, USA) according to the manufacturer's manual. The primers for the first round of PCR are: "S5-23R", TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNNNGTGGTACCTTTCT ATTCTCACTCT (forward), "N7-18", GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNTTCAACAGTTTC GGCCGAACCT (reverse). The "NNNNNNNN" denotes a 8-nt barcode sequence used for sample multiplexing. The primers for the second round of PCR are: "S510", AATGATACGGCGACCACCGAGATCTACACCGTCTAATTCGTCGGCAGCGTC (forward), "N710", CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGG (reverse). The volume of the 1st round of PCR was set as 50 μL with the phage DNA as the template, with 1×Q5 Buffer, 200 μM dNTP, 0.5 μM S5-23R forward primer, 0.5 μM N7-18, and 0.5 μL of Q5 polymerase. The reaction was then thermo-cycled: denatured at 98°C for 2 min; cycled as (98°C, 10 s; 55°C, 30 s; 72°C, 30 s) for 30 times; extended at 72°C for 2 min. PCR products were gel purified individually, and the concentration was measured using NanoDrop 2000c (Thermo Fisher, MA, USA). The volume of the 2nd round of PCR was set as 25 μL with 25 ng of the

purified first-round PCR product as the template, with 1×Q5 Buffer, 100 μM dNTP, 0.25 μM S510 forward primer, 0.25 μM N710 reverse primer, and 0.25 μL Q5 polymerase. The reaction was then thermo-cycled: denatured at 98°C for 2 min; cycled as (98°C, 10 s; 55°C, 30 s; 72°C, 30 s) for 10 times; extended at 72°C for 2 min. PCR products were gel purified individually. After concentration was measured, up to 100 barcoded samples with 100 ng of each were mixed together. The pooled library were sequenced on the Illumina HiSeq platform with paired-end 2×150 as the sequencing mode.

Bioinformatics analysis

The insert size of our library was 103 bp in total, which could be fully covered by both of the paired sequencing reads. To ensure high sequencing fidelity, the paired reads were required to give exactly the same insert sequence (*i.e.*, any mismatch in the 103-bp region between Read 1 and 2 excluded the insert sequence from downstream analysis). Within the insert sequence, there was an 8 bp N barcode, an 8 bp S barcode and a 36 bp varied DNA sequence. Each N and S barcode pair uniquely defined the serum sample where the insert sequence came from. The 36 bp DNA sequence, which corresponded to the peptide sequence displayed by phage, was translated into 12 aa peptide and assigned to the corresponding serum sample based on the barcode pair. After that, all peptides were counted and ranked in each serum sample. Finally, all count numbers were normalized by the total peptide count of each serum sample and Student's t-test was carried out for each peptide between the SLE and the healthy control group.

116 peptides significantly enriched in the SLE group (with Bonferroni adjusted *p* value <0.01 as the cut-off) were selected for further analysis.

The 116 selected peptides were used as the input feature of a multi-layer perceptron model from caret package (22) of the R language with 10-fold cross validation to select the most informative peptides that differentiate the SLE and healthy control group as much as possible. The ROC plot were generated by the ROCR (23) packages.

ELISA validation of the identified peptides

The C-terminal amidated synthetic model peptides were synthesized by GL Biochem Ltd. (Shanghai, China). Peptides were dissolved in DMSO and coupled with BSA to give a concentration of 1 mg/mL. High binding capacity 96-well ELISA plates (Corning, New York, USA) were coated with 50 μg/mL peptides in PBS, 100 μL/well, and kept sealed with parafilm at 4°C for overnight. Then washed twice with PBST, and blocked for 2 h at room temperature with blocking buffer (3% bovine serum albumin in PBST). The plates were washed twice with PBST. Each serum sample was diluted 1:100 in PBS buffer, added to the wells, and incubated at 37°C for 2 h. The plates were then washed six times with PBST, and excess buffer was aspirated after the last wash. Incubation with anti-human IgG-peroxidase (Sangon Biotech, Shanghai, China) was performed at 37 °C for 1 h. The plates were again washed six times with PBST, after excess buffer was aspirated, each well of the plates was added with 100 μL TMB (Tetramethylbenzidine) substrate (Sigma-Aldrich, Saint Louis, USA), and incubated at 37°C for 30 min. The reaction was stopped with 50 μL/well of 0.5 N H_2SO_4. Optical density was read

at 450 nm using a Behring EL311 ELISA microplate reader (Dade Behring Marburg Gmbh, Berlin, Germany). A typical four-parameter logistic nonlinear regression model was used for standard curve fitting for the ELISA, which was further used to estimate sample content (units/mL) from the absorbance measurement data.

Statistics analysis

All statistical analyses were performed using SPSS statistics 17.0, GraphPad Prism 6.0 and the R language. The mean values of un-transformed ELISA binding data were compared using Mann–Whitney U test. The Kruskal–Wallis test was used to compare multiple groups with Dunn's test as the post hoc test.

Results

Study design for the discovery of SLE specific peptides

This study was composed of two phases as follows: screening phase and validation phase (**Figure 1**). In the screening phase, we introduced an experimental approach using phage displayed peptide libraries and deep sequencing for the discovery of peptides that could be specifically recognized by serum IgG of SLE patients (**Figure 1A**). Briefly, screening was applied to enrich peptides that bind to antibodies in each serum sample. Protein G beads were used to capture the IgG bound peptides/phages, and the unbound phages were removed. The peptides enriched by each serum antibody repertoire were then sequenced using a barcoded amplicon library prepared from the separately enriched peptides/phages for each sample. To

control the experiments, a monoclonal anti-6×His antibody (repeated in multiple wells in a 96-well plate) was used as the positive control, PBS only and wild-type phage particles (not presenting any 12-aa peptides) were used as negative controls. The results of the anti-6×His antibody were consistent in the same 96-well plate, and from plate to plate (**Supplemental Figure S1A**). Using the online motif calculation tool MEME, the motif that was recognized by the anti-6×His antibody, *i.e.*, HHN, was identified (**Supplemental Figure S1B**). To assure reproducibility, 2 serum samples collected from healthy people (H7 and H9) were tested, and two repeats were carried out independently. The Pearson correlation coefficients of the raw peptide counts between the two repeats were 0.95 and 0.894 for H7 and H9 (**Supplemental Figure S1C and 1D**), respectively. These results indicated that satisfactory reproducibility could be achieved by our approach.

Following the establishment of the workflow, serum samples were then tested. Peptides bound by antibodies from individuals with SLE and healthy controls were analyzed, and peptides that showed significant binding differences between patients and healthy controls were identified. For the discovery phase and the follow-up validation phases, sera from 125 SLE patients and 130 healthy volunteers were included in this study (**Figure 1B, Tables I, II and III**). To test the specificity of the identified serum autoantibody biomarker for SLE diagnosis, sera from 51 people with other autoimmune-related diseases, *e.g.*, RA, DM, BD, and AS were also collected and tested.

Identification of serum autoantibody binding peptides in the screening phase

In the screening phase, matched sera from 50 SLE patients and 50 healthy controls were involved. Statistically, there was no age and gender bias between the SLE patient group and the healthy control group (**Table I**). A total of 100× of the phage displayed random peptide library (100×10^9) was applied to screen for specific peptides that were differentially recognized by autoantibodies in each serum sample. The enriched phages/peptides were identified using deep sequencing. We considered peptides with a *P*-value < 10^{-6} and fold change (SLE/Health) greater than 1.2 as potential biomarkers. This analysis identified 116 peptides that were highly reactive with the serum samples from SLE patients (**Figure 2A, Supplemental Table S1**). Computational analysis showed that some of 116 peptide candidates could significantly discriminate between the SLE samples and the control samples (P<0.001) (**Figure 2B**). We next calculated receiver-operating-characteristic (ROC) curves for all the 116 peptides (see Experimental Section). The results showed that the AUC values were in the range of 0.74-0.9. The AUC for one peptide was greater than 0.9, AUCs of 42.2% (49/116) were greater than 0.8, and 56.9% (66/116) were greater than 0.7 (**Figure 2C, Supplemental Table S2**).

Validation of the peptide candidates using ELISA

To test the applicability of the newly identified peptides for SLE diagnosis, an ELISA was developed, since ELISAs are commonly used in clinical settings. Ideally, it would be preferable to validate all 116 peptides using ELISA; however, in order to keep the cost and consumption of the samples under control, to begin with, we decided to validate 10 peptides with a higher differentiation capability (AUC values) (**Supplemental Table S3**). The peptides were

synthesized by a local company. The samples for ELISA validation included sera from an independent cohort of 60 SLE patients and 60 healthy controls with detailed clinical records (**Table II**). The 96-well plates were with individual synthesized peptides in each well and incubated with 100× diluted sera. The signals were read with an HRP-conjugated anti-human IgG. The results showed that four of the peptides had significant signals discriminating the SLE patients from the healthy controls (**Figure 3A**). To further assess the diagnostic accuracy of these peptides, an ROC analysis was also performed. The AUC values of the four peptides were greater than 0.7 among the peptides, three (SLE2018Val001/002/008) had AUC values greater than 0.8 (**Figure 3B, Supplemental Figure S2**). These results demonstrated that antibodies against these peptides have high potential to be used as biomarkers for SLE diagnosis.

Independent validation of the peptides with other autoimmune diseases

It is possible that the final panel of peptides may also react with the sera of other autoimmune diseases. To rule out this possibility, an additional set of sera from RA (n=7), DM (n=9), BD (n=20), and AS (n=15) was used to test the specificity of the biomarker candidates (**Table III**). The results showed that two of the four peptide candidates, *i.e.*, SLE2018Val001 and SLE2018Val002, have significant reaction signals with SLE patients as compared to that of other autoimmune diseases, such as RA, AS, and healthy controls (**Figure 4**). However, there was no significant difference between SLE with DM, SLE and BD diseases. In addition, for the other two peptides, *i.e.*, SLE2018Val006 and SLE2018Val008, there were significant differences between SLE patients and healthy controls, but not between SLE patients and

patients with other autoimmune diseases, except for a slight difference with AS patients for SLE2018Val006. However, the average values of these four peptides in SLE patient sera were generally higher than those in the sera from healthy controls and patients with other autoimmune diseases that we tested. The nonsignificant *P*-values may be due to the large variation among the limit number of samples that we managed to collect during the study period. Thus, a higher power of differentiation could be expected when an increased number of samples of other autoimmune-related diseases are tested. Nevertheless, these results indicated that at least two autoantibody biomarker candidates SLE2018Val001 and SLE2018Val002 are relatively specific for SLE.

Analysis of the combination of the four peptide candidates

As no single peptide identified could provide high specificity alone, the peptides were used in combination to improve the diagnostic accuracy. With the same ROC analysis used above, we plotted the sensitivity against 1-specificity in an ROC curve as a function of varying thresholds in a class prediction models (**Figure 5A**). Use of the four-peptide (SLE2018Val001/002/006/008) panel, named SLE2018Val, enabled discrimination of SLE patients from healthy controls with an AUC of 0.86 (sensitivity ~75%, specificity ~90%). These results showed that the candidate biomarker SLE2018Val has a higher specificity than the biomarkers used in routine SLE clinical testing, *i.e.,* an anti-DNA antibody with a specificity ~60% (24), or anti-Sm with a specificity of 40% (7).

Using the same method, we analyzed the candidate biomarker SLE2018Val aganist other autoimmune diseases. Similarly, SLE2018Val could discriminate SLE from other diseases, *i.e.,* the AUC of RA, BD, AS and DM was 0.97, 0.85, 0.96 and 0.88, respectively (**Figure 5B**). All these results demonstrated that the panel of four peptides (SLE2018Val) has the potential to be developed into a clinical biomarker test for SLE diagnostics.

Discussion

In this study, we established a general strategy for discovering serum autoantibody biomarkers. The key components of this strategy are the vast diversity of the phage display random peptide library and the extraordinary sequencing power of deep sequencing. By taking SLE as an example, we identified a set of 116 autoantibodies against specific peptides with the capability of differentiating SLE patients from healthy people. After several rounds of validation, a panel of four peptides (SLE2018Val001, SLE2018Val002, SLE2018Val006, SLE2018Val008) were determined as high potential SLE biomarkers, either individually or in combination (SLE2018Val) for the diagnosis of SLE. The results of the remaining six peptides were shown in **Figure S2**. The AUC values were lower than those of the four-peptide panel.

In the screening phase, a large amount of the peptides found with statistically significant antibody reactions were unique to each sample, indicating that in general, each person may possess a unique autoantibody fingerprint. In this case, the impact of phenotype and disease history on autoimmunity may need to be explored. Furthermore, screening a large number of individuals of different genetic background may additionally uncover the correlations between

autoreactivity and SLE disease phenotypes, antibody variable domain alleles, and other immunogenetic modifiers.

Our study is unique compared with previous studies on the discovery of SLE biomarkers. Traditionally, researchers have applied microarrays of whole antigen (protein or lipid) to characterize the sera of autoimmune diseases (21, 25-28). However, these arrays are limited in that they only contain known autoantigens. The use of phage displayed random peptide library allows the unbiased assessment of antibody binding to classify patients. Unlike protein microarrays (29, 30), the display of random peptides is not restricted to proteins that have been cloned and can be expressed recombinantly. Moreover, for studies involving a large number of samples, the cost per-sample of our strategy is roughly two orders of magnitude less than that of microarray-based alternatives (31). In this study, we screened serum autoantibodies using a phage displayed random peptide library, which contained ~1×10^9 different peptides, increasing our chance of identifying potential biomarkers in an unbiased manner.

To our knowledge, the first attempt to use a displayed random peptide library and deep sequencing for autoantibody based serum biomarker discovery was by Pantazes *et al*. (32), who used an *E. coli*-displayed peptide library to identify celiac disease biomarkers. Pantazes *et al.* identified peptides that showed increased binding to antibodies in the patient sera. Then, they claimed that motifs were successfully calculated from these peptides. However, much greater diversity could be achieved by phage display than bacterial display. In addition, high quality phage display library is commercially available, while bacteria display library need to be custom prepared by researchers. Greater diversity means greater coverage of possible

epitope/mimotopes, which may explain why we identified 116 peptides as potential biomarkers in the discovery phase. However, even with these many peptides, we failed to identify any meaningful common motifs. One plausible explanation is that the autoimmune response of SLE is complicated and mediated by many antigens. In addition, no specific human antigen has been found complete matching with our validated peptides (**Supplemental Table S4**) using BLASTP, which indicated that the components causing an immune response may not necessarily belong to proteins; thus, none of the autoantibodies are dominant in the population; making it difficult to identify significantly enriched motifs.

In this study, we developed a standard protocol for serum autoantibody biomarker screening with the phage displayed random peptide library. To ensure the reliability of the comparison between the patients and healthy controls, the concentration of total IgG in the sera was measured by ELISA. In another ongoing study, using a similar phage enrichment and deep sequencing workflow, we could readily enrich expected peptides/phages for antibodies, for example, histidine rich peptides for an anti-His antibody. To control the entire experimental procedure, a set of widely applied antibodies, *i.e.*, anti-His, anti-Flag, and anti-GST, were included. An optimal pipeline for the bioinformatics analysis of deep sequencing data has also been established.

Our results are consistent with a range of clinical features, including anti-dsDNA (24), anti-Sm and anti-RNP (7), which have sensitivities and specificities ranging from 60% to 80% in discriminating SLE patients and healthy controls. The combination of the four peptides (SLE2018Val) had sensitivity and specificity of 0.75 and 0.90, respectively, which is equal to

or better than many of the known SLE biomarkers. The peptide set may be useful in combination with routinely used autoantibody assays, for example, anti-dsDNA screening.

Though the peptides that we identified in this study could serve as potential biomarkers for SLE diagnosis. It is necessary to point out that more analysis and validation is required before we move the biomarkers to the clinics. First, the analysis of a larger number of patient samples derived from a more diverse population will be necessary. Second, it will be important to test samples collected from patients with different SLE disease activity index (SLEDAI) scores to disease activity (DA) scores because early detection of developing disease is an important clinical goal (33). Since SLE is highly heterogeneous, not all manifestations are included in the SLEDAI, making reliable patient assessment challenging. Thus, it will be interesting to see the differences between SLE pathogenesis and DA for biomarker purposes. Additionally, it will be interesting to identify subgroup specific serum biomarkers for SLE in future studies. Third, for practical reasons, we validated only 10 of the peptides. It is highly possible that peptides of equal or even better performance exist in the remaining 106 peptides. Ideally, the best approach is to unbiasedly validate all 116 peptides in a more efficient and economical manner, for example, peptide microarray, rather than ELISA. Finally, if these biomarkers are indeed validated, then it will be of great interest to identify the native antigens that the peptides represent or mimic. There are two applicable approaches for antigen identification, one method is to blast the peptide against the sequence of the human proteome, the other is to pull down the corresponding IgGs with the peptides, and then use the IgGs to enrich the antigens for

identification. To facilitate the identification of the autoantigens, a phage displayed library covering the whole human proteome is an alternative choice (31).

While this study focused on SLE, our approach should be generally applicable to other autoimmune or autoimmune related diseases, such as rheumatoid arthritis, inflammatory bowel disease, or inflammatory myopathies. Our strategy may also be suitable for identifying biomarkers to predict responses to targeted therapies and reducing the stratification time.

We will apply our strategy to a series of other diseases and construct a database to include all the data from these studies. As the size of the database increases, since deep sequencing data are digital and can generally be compared among many diseases in the database, we expect the discovery of rare yet significantly disease-associated biomarkers/autoantibodies for some of the diseases with high disease specificity to increase, without the necessity of repeating many of the experiments.

Phage displayed random peptide library is a complementary resource for other species specific peptidomes in the discovery of serum biomarkers. A synthetic human peptidome has been applied to a variety of autoimmune diseases, *i.e.*, paraneoplastic neurological syndromes (17), multiple sclerosis, type 1 diabetes and rheumatoid arthritis (34). A synthetic human viruses peptidome (Virscan) was used to comprehensively analyze antiviral antibodies in human sera (35). Our approach is similar to PhIP-seq. The major difference between our technology and PhIP-seq is the phage library. Usually, a species-specific phage displayed peptide library of high coverage is synthesized and used in PhIP-seq, *e.g.*, human and viruses. Thus, PhIP-seq is well-suited for protein based auto-antigen (without any modification and

mutation) discovery, and the results could be clearly explained in a biologically meaningful context. However, PhIP-seq may fail to identify antigens that are mutated proteins, proteins with aberrant post-translational modifications, or not protein based. In our approach, a phage displayed random peptide library (Ph.D.12) was applied. This library has a diversity of 10^9, which is $10^4 \sim 10^5$ times larger than the proteome specific peptide library used in PhIP-seq. With this diversity, it is possible that the random peptide library could cover a large enough binding space, *i.e.*, for any of the antigens (protein based or not protein based), that peptides that mimic the antigen of interest are present. Therefore, no matter what the original antigen is, peptides that could differentiate patients and healthy people could be identified. In addition, a high quality random peptide library could be directly purchased at relatively low cost. Thus, for biomarker identification, we think our approach is more generally applicable and accessible than traditional PhIP-seq.

In summary, we identified a panel of four individual and combined (SLE2018Val) peptides as high-potential serum biomarkers for SLE. These autoantibody-based biomarkers could differentiate SLE patients from healthy individuals with a high degree of sensitivity and specificity, either individually or in a combination. Our serum biomarker discovery strategy is generally applicable and could be easily applied to other diseases; thus, a large set of biomarkers for a variety of diseases could be identified in the near future.

Acknowledgements

This study was supported in part by The National Key Research and Development Program of China Grant 2016YFA0500600, National Natural Science Foundation of China Grants 31670831, 31370813, and 31501054. This study was also supported by the key Research and Development Program of Yantai Grant 2019XDHZ089, Shandong Provincial Natural Science Foundation of China Grant ZR2019PC052.

Author Contributions

SCT conceived and designed the project. HL carried out most of the computational analyses. HQ provided key reagents. FLW and DYL performed most of the experiments. HHD, YJT and NS provided the serum samples. ZWX, MLM and SJG contributed reagents or provided laboratory assistance. FLW and SCT interpreted results and wrote the manuscript. JFW and XDZ commented on the manuscript.

Competing interests

The authors declare that they have no conflict of interests.

References

1. Luijten, K. M. A. C., Tekstra, J., Bijlsma, J. W. J., and Bijl, M. (2012) The Systemic Lupus Erythematosus Responder Index (SRI); A new SLE disease activity assessment. *Autoimmun Rev* 11, 326-329

2. Danchenko, N., Satia, J. A., and Anthony, M. S. (2006) Epidemiology of systemic lupus erythematosus: a comparison of worldwide disease burden. *Lupus* 15, 308-318

3. Beeson, P. B. (1994) Age and sex associations of 40 autoimmune diseases. *The American journal of medicine* 96, 457-462

4. Bernatsky, S., Boivin, J. F., Joseph, L., Manzi, S., Ginzler, E., Gladman, D. D., Urowitz, M., Fortin, P. R., Petri, M., Barr, S., Gordon, C., Bae, S. C., Isenberg, D., Zoma, A., Aranow, C., Dooley, M. A., Nived, O.,

Sturfelt, G., Steinsson, K., Alarcon, G., Senecal, J. L., Zummer, M., Hanly, J., Ensworth, S., Pope, J., Edworthy, S., Rahman, A., Sibley, J., El-Gabalawy, H., McCarthy, T., St Pierre, Y., Clarke, A., and Ramsey-Goldman, R. (2006) Mortality in systemic lupus erythematosus. *Arthritis and rheumatism* 54, 2550-2557

5. Yaniv, G., Twig, G., Shor, D. B., Furer, A., Sherer, Y., Mozes, O., Komisar, O., Slonimsky, E., Kiang, E., Lotan, E., Welt, M., Maraj, I., Shina, A., Amital, H., and Shoenfeld, Y. (2015) A volcanic explosion of autoantibodies in systemic lupus erythematosus: A diversity of 180 different antibodies found in SLE patients. *Autoimmun Rev* 14, 75-79

6. Solomon, D. H., Kavanaugh, A. J., Schur, P. H., and American College of Rheumatology Ad Hoc Committee on Immunologic Testing, G. (2002) Evidence-based guidelines for the use of immunologic tests: antinuclear antibody testing. *Arthritis and rheumatism* 47, 434-444

7. Benito-Garcia, E., Schur, P. H., Lahita, R., and American College of Rheumatology Ad Hoc Committee on Immunologic Testing, G. (2004) Guidelines for immunologic laboratory testing in the rheumatic diseases: anti-Sm and anti-RNP antibody tests. *Arthritis and rheumatism* 51, 1030-1044

8. Quan, J., Lakhanpal, A., Reddy, M. M., Zaman, S., Li, Q. Z., German, D. C., Olsen, N. J., Kodadek, T., and Karp, D. R. (2014) Discovery of biomarkers for systemic lupus erythematosus using a library of synthetic autoantigen surrogates. *Journal of immunological methods* 402, 23-34

9. Reddy, M. M., Wilson, R., Wilson, J., Connell, S., Gocke, A., Hynan, L., German, D., and Kodadek, T. (2011) Identification of candidate IgG biomarkers for Alzheimer's disease via combinatorial library screening. *Cell* 144, 132-142

10. Bruzelius, M., Iglesias, M. J., Hong, M. G., Sanchez-Rivera, L., Gyorgy, B., Souto, J. C., Franberg, M., Fredolini, C., Strawbridge, R. J., Holmstrom, M., Hamsten, A., Uhlen, M., Silveira, A., Soria, J. M., Smadja, D. M., Butler, L. M., Schwenk, J. M., Morange, P. E., Tregouet, D. A., and Odeberg, J. (2016) PDGFB, a new candidate plasma biomarker for venous thromboembolism: results from the VEREMA affinity proteomics study. *Blood* 128, e59-e66

11. Wang, X., Yu, J., Sreekumar, A., Varambally, S., Shen, R., Giacherio, D., Mehra, R., Montie, J. E., Pienta, K. J., Sanda, M. G., Kantoff, P. W., Rubin, M. A., Wei, J. T., Ghosh, D., and Chinnaiyan, A. M. (2005) Autoantibody signatures in prostate cancer. *The New England journal of medicine* 353, 1224-1235

12. Hecker, M., Fitzner, B., Wendt, M., Lorenz, P., Flechtner, K., Steinbeck, F., Schroder, I., Thiesen, H. J., and Zettl, U. K. (2016) High-Density Peptide Microarray Analysis of IgG Autoantibody Reactivities in Serum and Cerebrospinal Fluid of Multiple Sclerosis Patients. *Molecular & Cellular Proteomics* 15, 1360-1380

13. Zhu, H., Luo, H., Yan, M., Zuo, X., and Li, Q. Z. (2015) Autoantigen Microarray for High-throughput Autoantibody Profiling in Systemic Lupus Erythematosus. *Genomics, proteomics & bioinformatics* 13, 210-218

14. Kanter, J. L., Narayana, S., Ho, P. P., Catz, I., Warren, K. G., Sobel, R. A., Steinman, L., and Robinson, W. H. (2006) Lipid microarrays identify key mediators of autoimmune brain inflammation. *Nature medicine* 12, 138-143

15. Lueking, A., Possling, A., Huber, O., Beveridge, A., Horn, M., Eickhoff, H., Schuchardt, J., Lehrach, H., and Cahill, D. J. (2003) A nonredundant human protein chip for antibody screening and serum profiling. *Molecular & cellular proteomics : MCP* 2, 1342-1349

16. Robinson, W. H., DiGennaro, C., Hueber, W., Haab, B. B., Kamachi, M., Dean, E. J., Fournel, S., Fong, D., Genovese, M. C., de Vegvar, H. E., Skriner, K., Hirschberg, D. L., Morris, R. I., Muller, S., Pruijn, G. J.,

van Venrooij, W. J., Smolen, J. S., Brown, P. O., Steinman, L., and Utz, P. J. (2002) Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nature medicine* 8, 295-301

17. Larman, H. B., Zhao, Z., Laserson, U., Li, M. Z., Ciccia, A., Gakidis, M. A., Church, G. M., Kesari, S., Leproust, E. M., Solimini, N. L., and Elledge, S. J. (2011) Autoantigen discovery with a synthetic human peptidome. *Nature biotechnology* 29, 535-541

18. Christiansen, A., Kringelum, J. V., Hansen, C. S., Bogh, K. L., Sullivan, E., Patel, J., Rigby, N. M., Eiwegger, T., Szepfalusi, Z., de Masi, F., Nielsen, M., Lund, O., and Dufva, M. (2015) High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum. *Scientific reports* 5, 12913

19. Rao, S. S., Mohan, K. V. K., Nguyen, N., Abraham, B., Abdouleva, G., Zhang, P., and Atreya, C. D. (2010) Peptides panned from a phage-displayed random peptide library are useful for the detection of Bacillus anthracis surrogates B. cereus 4342 and B. anthracis Sterne. *Biochem Bioph Res Co* 395, 93-98

20. Hochberg, M. C. (1997) Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis and rheumatism* 40, 1725-1725

21. Yang, L., Wang, J., Li, J., Zhang, H., Guo, S., Yan, M., Zhu, Z., Lan, B., Ding, Y., Xu, M., Li, W., Gu, X., Qi, C., Zhu, H., Shao, Z., Liu, B., and Tao, S. C. (2016) Identification of Serum Biomarkers for Gastric Cancer Diagnosis Using a Human Proteome Microarray. *Molecular & cellular proteomics : MCP* 15, 614-623

22. Turner, H. (2011) A peer-reviewed, open-access publication of the R Foundation for Statistical Computing. *R J* 3, 3-3

23. Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics* 21, 3940-3941

24. Kavanaugh, A. F., Solomon, D. H., and American College of Rheumatology Ad Hoc Committee on Immunologic Testing, G. (2002) Guidelines for immunologic laboratory testing in the rheumatic diseases: anti-DNA antibody tests. *Arthritis and rheumatism* 47, 546-555

25. Li, Q. Z., Karp, D. R., Quan, J., Branch, V. K., Zhou, J., Lian, Y., Chong, B. F., Wakeland, E. K., and Olsen, N. J. (2011) Risk factors for ANA positivity in healthy persons. *Arthritis research & therapy* 13, R38

26. Li, Q. Z., Zhou, J., Wandstrat, A. E., Carr-Johnson, F., Branch, V., Karp, D. R., Mohan, C., Wakeland, E. K., and Olsen, N. J. (2007) Protein array autoantibody profiles for insights into systemic lupus erythematosus and incomplete lupus syndromes. *Clinical and experimental immunology* 147, 60-70

27. Yonekubo, Y., Wu, P., Esechie, A., Zhang, Y. E. Q., and Du, G. W. (2010) Characterization of new serum biomarkers in breast cancer using lipid microarrays. *Tumor Biol* 31, 181-187

28. Wandstrat, A. E., Carr-Johnson, F., Branch, V., Gray, H., Fairhurst, A. M., Reimold, A., Karp, D., Wakeland, E. K., and Olsen, N. J. (2006) Autoantibody profiling to identify individuals at risk for systemic lupus erythematosus. *Journal of autoimmunity* 27, 153-160

29. Yu, X., Song, L., Petritis, B., Bian, X., Wang, H., Viloria, J., Park, J., Bui, H., Li, H., Wang, J., Liu, L., Yang, L., Duan, H., McMurray, D. N., Achkar, J. M., Magee, M., Qiu, J., and LaBaer, J. (2017) Multiplexed Nucleic Acid Programmable Protein Arrays. *Theranostics* 7, 4057-4070

30. Atak, A., Mukherjee, S., Jain, R., Gupta, S., Singh, V. A., Gahoi, N., K, P. M., and Srivastava, S. (2016) Protein microarray applications: Autoantibody detection and posttranslational modification. *Proteomics* 16, 2557-2569

31. Mohan, D., Wansley, D. L., Sie, B. M., Noon, M. S., Baer, A. N., Laserson, U., and Larman, H. B. (2018) PhIP-Seq characterization of serum antibodies using oligonucleotide-encoded peptidomes. *Nature protocols* 13, 1958-1978

32. Pantazes, R. J., Reifert, J., Bozekowski, J., Ibsen, K. N., Murray, J. A., and Daugherty, P. S. (2016) Identification of disease-specific motifs in the antibody specificity repertoire via next-generation sequencing. *Scientific reports* 6, 30312

33. Bombardier, C., Gladman, D. D., Urowitz, M. B., Caron, D., and Chang, C. H. (1992) Derivation of the SLEDAI. A disease activity index for lupus patients. The Committee on Prognosis Studies in SLE. *Arthritis and rheumatism* 35, 630-640

34. Larman, H. B., Laserson, U., Querol, L., Verhaeghen, K., Solimini, N. L., Xu, G. J., Klarenbeek, P. L., Church, G. M., Hafler, D. A., Plenge, R. M., Nigrovic, P. A., De Jager, P. L., Weets, I., Martens, G. A., O'Connor, K. C., and Elledge, S. J. (2013) PhIP-Seq characterization of autoantibodies from patients with multiple sclerosis, type 1 diabetes and rheumatoid arthritis. *Journal of autoimmunity* 43, 1-9

35. Xu, G. J., Kula, T., Xu, Q., Li, M. Z., Vernon, S. D., Ndung'u, T., Ruxrungtham, K., Sanchez, J., Brander, C., Chung, R. T., O'Connor, K. C., Walker, B., Larman, H. B., and Elledge, S. J. (2015) Viral immunology. Comprehensive serological profiling of human populations using a synthetic human virome. *Science* 348, aaa0698

Figure legends

Figure 1. Schematic diagram and workflow. A. The schematic for phage display screening and next generation sequencing. **B**. This study is composed of three major phases, *i.e.*, screening phase, validation phase I, and validation phase II.

Figure 2. The candidate peptide biomarkers were discovered in the screening phase. A. The peptides which could differentiate SLE patients from healthy controls were discovered using next generation sequencing. **B**. Reactivity of the candidate peptides with SLE patients and health controls. **C**. Receiver operating characteristic curve analysis of the candidate peptides between SLE patients and health controls.

Figure 3. Four candidates of SLE serum biomarker were confirmed in validation phase. A. Reactivity of the serum biomarkers with sera of SLE patients and healthy controls in the ELISA validation. **B**. Receiver operating characteristic curve analysis of the four candidate peptides with SLE patients and health controls.

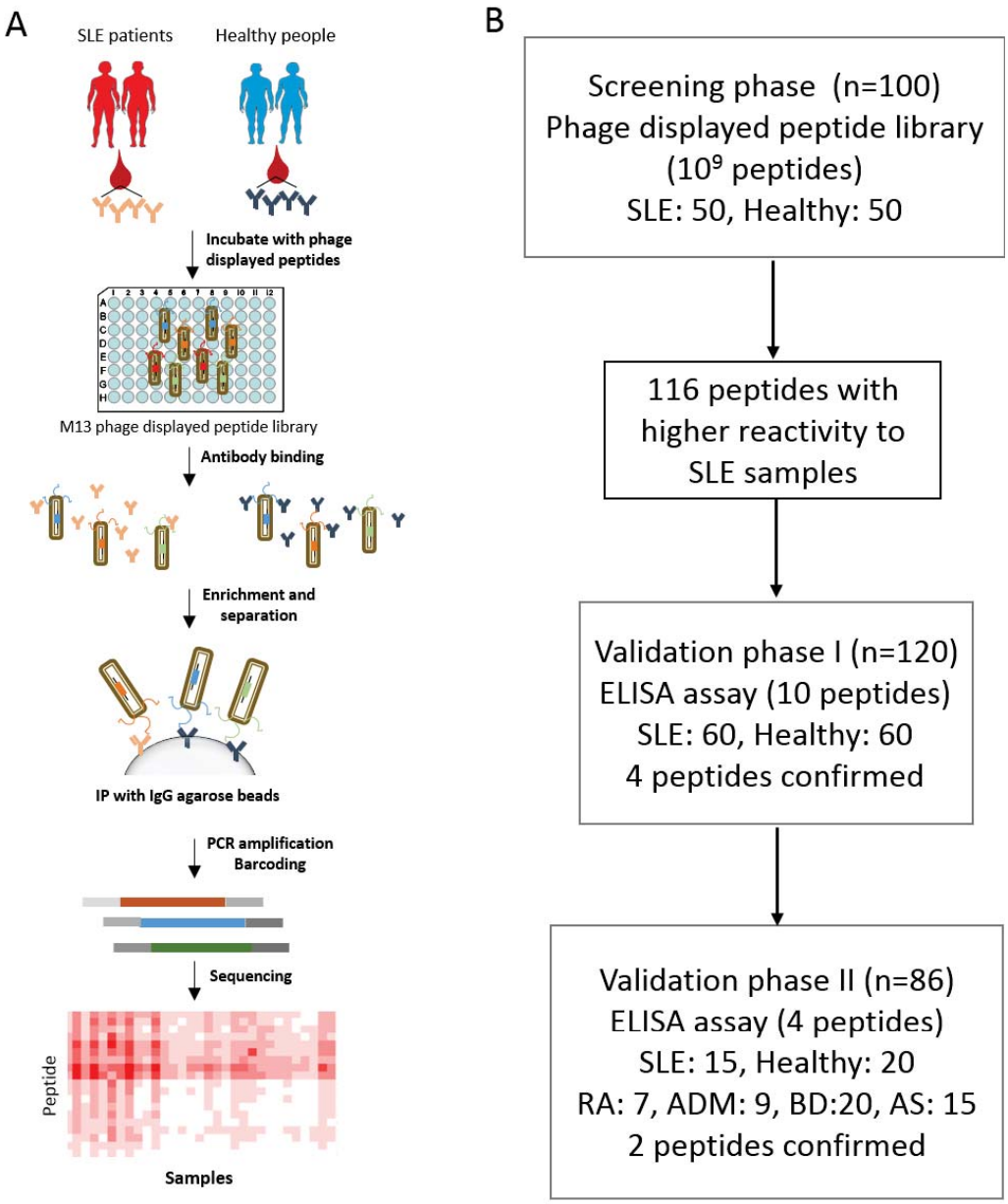
Figure 4. Assessment of the specificity of the four biomarkers with other related autoimmune diseases. Comparison of the SLE patients, healthy controls and other autoimmune patients, *i.e.*, RA, ADM, BD and AS. Asterisks indicate statistical difference as compared to the SLE group ($p < 0.05$).

Figure 5. Receiver operating characteristic curve analysis using the combinational panel of the four biomarkers. A. The best model and the classifier for SLE against healthy controls. Based on the cross validation, a best model [(3.479×expression level of SLE2018Val001) + (13.131×expression level of SLE2018Val002) - (0.051×expression level of SLE2018Val006) + (7.517×expression level of SLE2018Val008) - 10.263] was generated with the final 4-peptide panel using generalized linear model from the R language. **B.** The best model and the classifier for SLE to RA, BD, AS and DM.

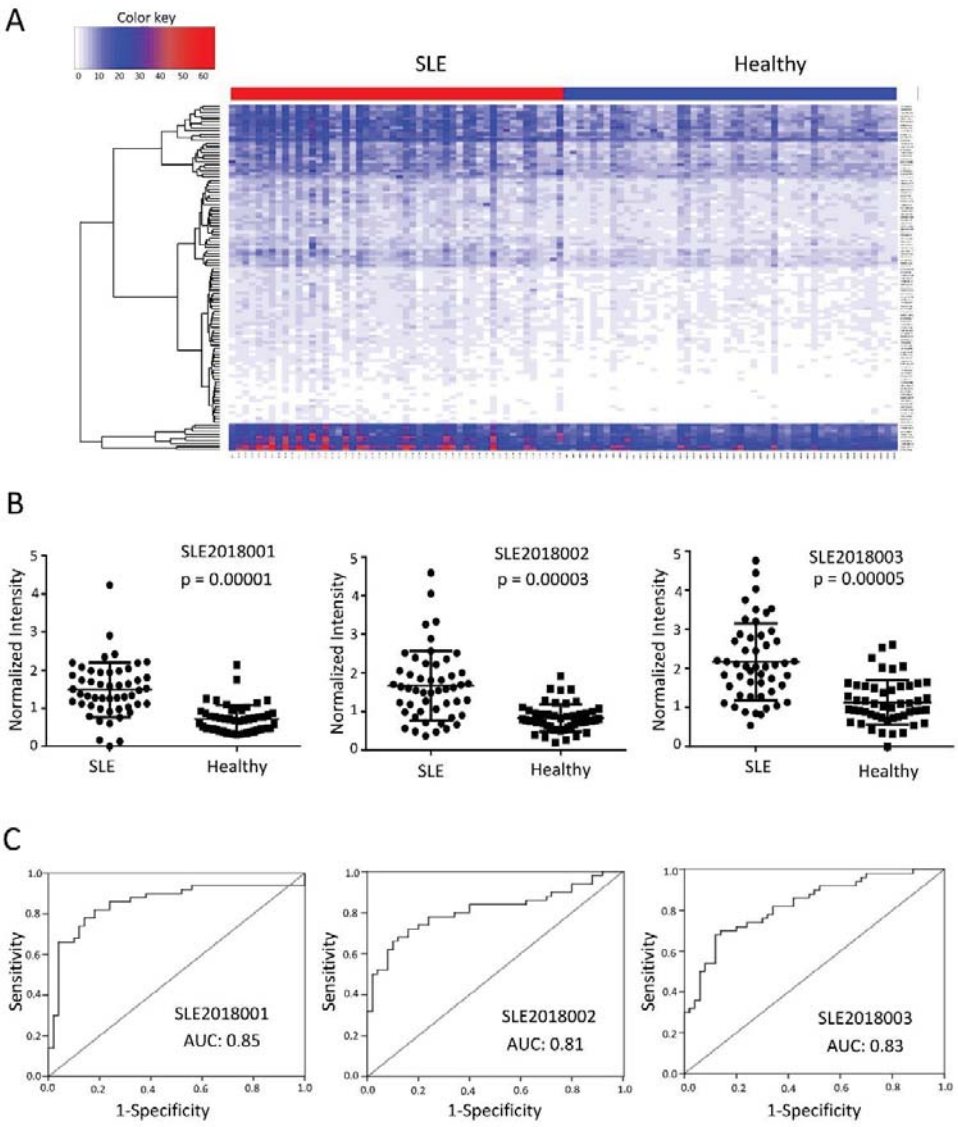
Table I. Characteristics of study participants in the screening phase.

Table II. Characteristics of study participants in the validation phase-I.

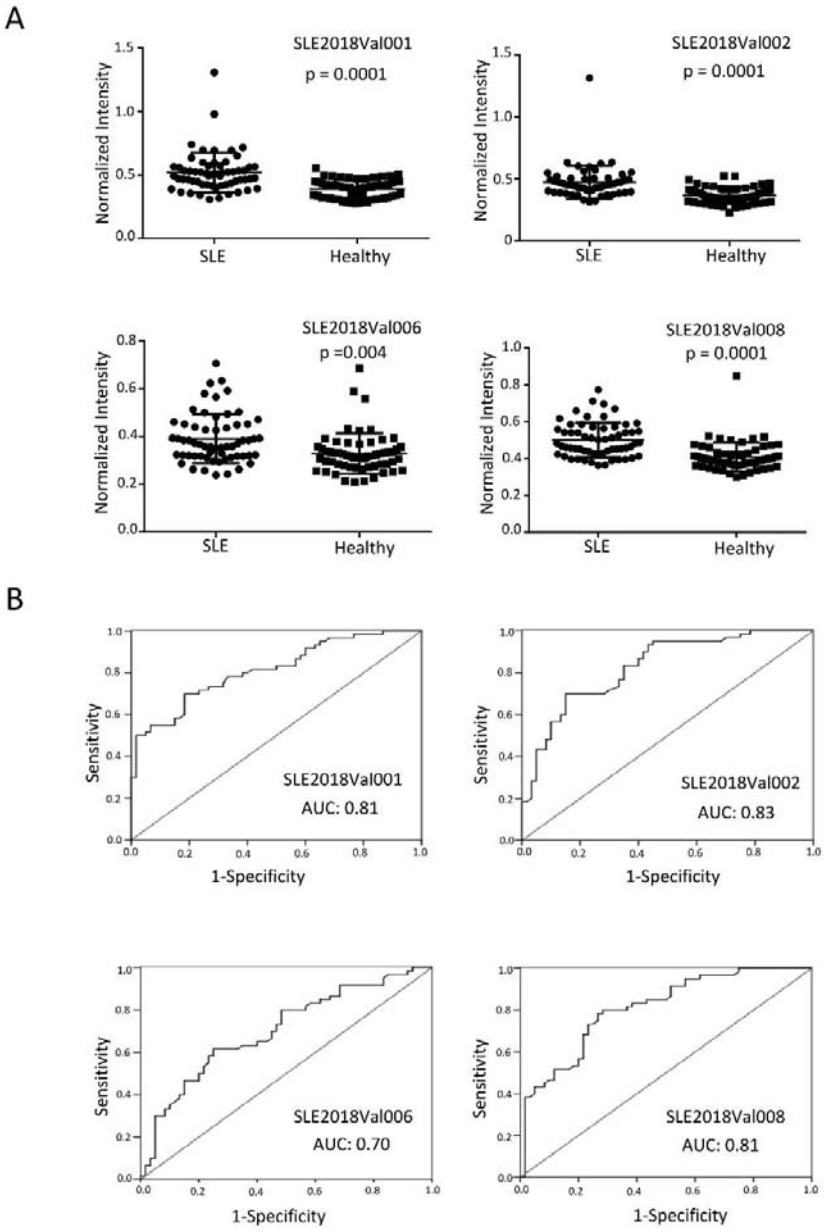
Table III. Characteristics of study participants in the validation phase-II.



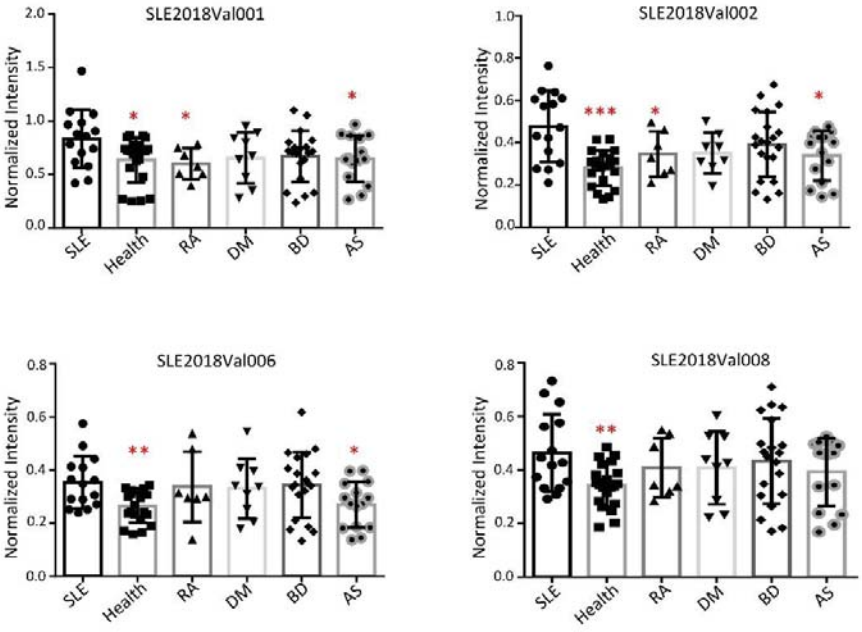
Wu *et. al.*, Figure 1



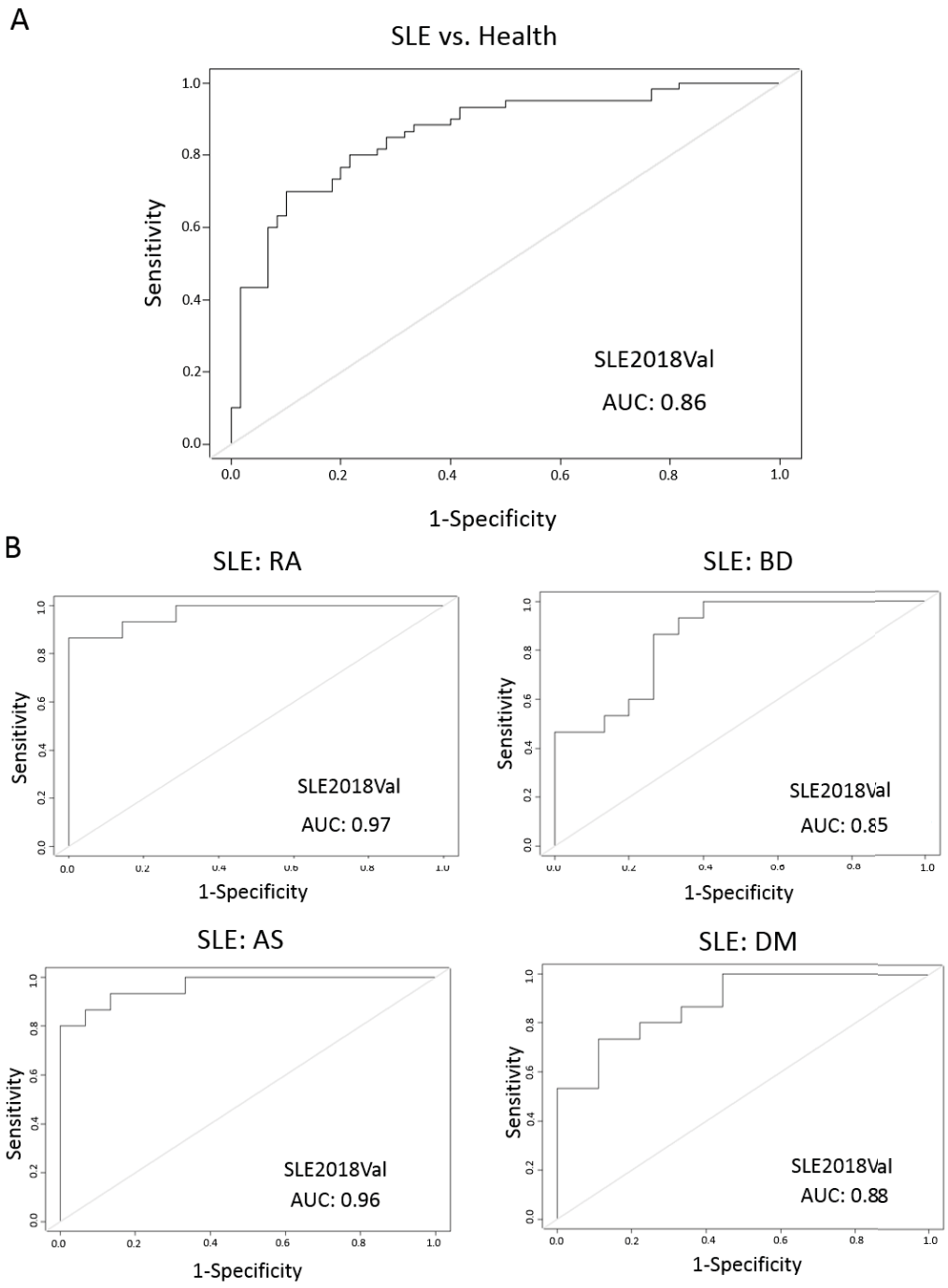
Wu *et. al.*, Figure 2



Wu *et. al.*, Figure 3



Wu *et. al.*, Figure 4



Wu *et. al.*, Figure 5

Table I

Characteristics of study participants in the screening phase

Variable	SLE count (n=50)			Health count (n=50)		
	No.	Mean	%	No.	Mean	%
Age, years						
Mean		37.8			38.1	
Interquartile		30-45			23-53	
Sex						
Male	5		10	5		10
Female	45		90	45		90
Disease Duration						
Median		4.00				
Interquartile		1.00-8.96				
SLEDAI						
Median		7				
Interquartile		5-13				
Hypocomplementemia			77.5			
Anti-dsDNA positive			38.7			
Antinuclear antibody positive			44.0			
Clinical Manifestation						
Fever			41.9			
Malar Rash			34.9			
Arthralgia			27.9			
Pulmonary Artery Hypertension			7.0			
Lupus Nephritis			67.4			
Leukocytopenia			27.9			
Thrombocytopenia			11.6			
Sjogren Syndrome			7.0			
Concurrent Medicine Use						
Glucocorticoids			88.4			
Hydrocloriquine			39.5			
Mycophenolate mofetil			11.6			
Cyclosporin			0.0			
Tacrolimus			4.7			
Cyclophosphamide			2.3			
Aspirin			18.6			
Calcitrol			76.7			

SLEDAI : Systemic Lupus Erythematosus Disease Activity Index

Table II

Characteristics of study participants in the validation phase-I

Variable	SLE count (n=60)			Health count (n=60)		
	No.	Mean	%	No.	Mean	%
Age, years						
Mean		41.4			41.2	
Interquartile		30-51			23-57	
Sex						
Male	9		15	10		16
Female	51		85	50		84
Disease Duration						
Median		0.26				
Interquartile		0.08-1.03				
SLEDAI						
Median		7				
Interquartile		4-10				
Hypocomplementemia			78.0			
Anti-dsDNA positive			50.0			
Antinuclear antibody positive			16.0			
Clinical Manifestation						
Fever			10.0			
Malar Rash			15.0			
Arthralgia			21.7			
Pulmonary Artery Hypertension			1.7			
Lupus Nephritis			65.0			
Leukocytopenia			38.3			
Thrombocytopenia			13.3			
Sjogren Syndrome			5.0			
Concurrent Medicine Use						
Glucocorticoids			84.3			
Hydrocloriquine			52.9			
Mycophenolate mofetil			7.8			
Cyclosporin			3.9			
Tacrolimus			2.0			
Cyclophosphamide			5.9			
Aspirin			23.5			
Calcitrol			86.3			

SLEDAI : Systemic Lupus Erythematosus Disease Activity Index

Table III

Characteristics of study participants in the validation phase-II

Variable	No.	Mean	%
SLE	15		
Age, years		35.9	
Male	2		13
Female	13		87
RA	7		
Age, years		53.2	
Male	2		28
Female	5		72
DM	9		
Age, years		51.0	
Male	3		33
Female	6		66
BD	20		
Age, years		34.3	
Male	9		21
Female	11		79
AS	15		
Age, years		36.1	
Male	11		73
Female	4		27
Health	20		
Age, years		40.8	
Male	0		0
Female	20		100

The abbreviations used are as follow: RA, rheumatoid arthritis ; DM, dermatomyositis; BD, behcet disease; AS, ankylosing spondylitis