

Severe acute respiratory syndrome diagnostics using a coronavirus protein microarray

Heng Zhu^{*†‡§}, Shaohui Hu^{†‡}, Ghil Jona^{*‡}, Xiaowei Zhu^{†¶}, Nate Kreiswirth^{||}, Barbara M. Willey^{||}, Tony Mazzulli^{||}, Guozhen Liu^{†**}, Qifeng Song[†], Peng Chen[†], Mark Cameron^{||}, Andrea Tyler^{||}, Jian Wang[†], Jie Wen[†], Weijun Chen[†], Susan Compton^{††}, and Michael Snyder^{*¶†‡‡}

Departments of *Molecular, Cellular, and Developmental Biology and ††Comparative Medicine, and †¶Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520; †Biochip Platform Division, Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 101300, China; †Department of Microbiology, Mount Sinai Hospital, Toronto, ON, Canada M5G 1X5; and **College of Life Sciences, Agricultural University of Hebei, Hebei, Baoding 071001, China

Communicated by Dieter Söll, Yale University, New Haven, CT, January 16, 2006 (received for review October 10, 2005)

To monitor severe acute respiratory syndrome (SARS) infection, a coronavirus protein microarray that harbors proteins from SARS coronavirus (SARS-CoV) and five additional coronaviruses was constructed. These microarrays were used to screen ≈400 Canadian sera from the SARS outbreak, including samples from confirmed SARS-CoV cases, respiratory illness patients, and healthcare professionals. A computer algorithm that uses multiple classifiers to predict samples from SARS patients was developed and used to predict 206 sera from Chinese fever patients. The test assigned patients into two distinct groups: those with antibodies to SARS-CoV and those without. The microarray also identified patients with sera reactive against other coronavirus proteins. Our results correlated well with an indirect immunofluorescence test and demonstrated that viral infection can be monitored for many months after infection. We show that protein microarrays can serve as a rapid, sensitive, and simple tool for large-scale identification of viral-specific antibodies in sera.

infectious disease | protein chip | virus diagnostics

In November 2002, an outbreak of severe acute respiratory syndrome (SARS) occurred in southern China and rapidly spread across five continents. SARS was characterized by fever and respiratory compromise; the World Health Organization estimated that SARS infected 8,439 individuals with a mortality rate of ≈9% overall and 40% in people older than 60 years (1). A novel coronavirus, SARS coronavirus (SARS-CoV), was identified as the etiological agent for the illness and was found to be related to, but distinct from, other coronaviruses, including two previously identified human coronaviruses, HCoV-229E and HCoV-OC43, single-stranded RNA viruses that collectively cause ≈30% of common colds in humans (2). Like other coronaviruses, SARS-CoV encodes two RNA-dependent replicases, 1a and 1b, a spike protein, a small envelope protein, a membrane protein, and a nucleocapsid (N) protein, as well as nine predicted proteins that lack significant similarity to any known proteins.

At present, no effective treatment of SARS is available. Isolation and stringent infection-control practices were the sole means to control the epidemic. Hence, rapid, accurate, and early diagnostic tests are necessary to monitor the course of the disease.

The World Health Organization classification for SARS infection in adults is based on four criteria: fever, respiratory symptoms, close proximity to infected individuals, and radiological evidence of lung infiltrates (3). Several diagnostic approaches have also been used for detecting SARS-CoV, including RT-PCR techniques, ELISAs, and the indirect immunofluorescence test (IIFT). RT-PCR is sensitive, specific, and useful during the infection (4–7). However, it is not useful once the infection is cleared and can be challenging to implement in clinical application; the collection of samples such as nasopharyngeal or bronchial alveolar aspirates from SARS patients is dangerous and can put healthcare workers at high risk. ELISAs tend not to be highly sensitive and usually

require large amounts of sample (8–11). Moreover, existing ELISAs, such as one manufactured by Euroimmun (Luebeck, Germany), use whole viral extracts, thereby increasing the chance of misdiagnosis due to crossreactivity with proteins from other viruses. Currently, an IIFT kit (Euroimmun) to detect SARS IgG antibody response is considered the serological gold-standard method in the clinic. However, IIFT limitations include (i) difficulty in diagnosis in the urgent acute phases of the disease, (ii) failure to diagnose ≈5% of sera that contain high concentrations of antinuclear factor, and (iii) visual inspection of fluorescently stained cells, which is both subjective and of modest throughput. Thus, more tests for diagnosing the disease need to be developed.

We report the construction of a coronavirus proteome microarray that contains the entire proteomes of the human SARS-CoV and HCoV-229E viruses and the partial proteomes of human HCoV-OC43, mouse MHVA59, bovine coronavirus (BCoV), and feline coronavirus (FIPV). The coronavirus protein microarrays were used to screen serum samples collected from fever and respiratory patients during the period of SARS outbreak in Beijing and Toronto. Algorithms to optimally diagnose SARS-infected patients were devised to generate a microarray test that is rapid, sensitive, accurate, and adaptable for detection of many other types of viral infections.

Results

Development of a Coronavirus Protein Microarray and a SARS Detection Assay. A protein microarray approach was developed to rapidly identify SARS-CoV and other coronavirus-infected patients with high sensitivity and accuracy. Genes or gene fragments that cover the entire genome of SARS-CoV and the majority of the HCoV-229E and MHVA59 genomes were amplified by PCR and cloned into a yeast expression vector that produces the viral proteins with GST at their N terminus (Fig. 1). Using the limited sequence information available at the time, regions of the BCoV, HCoV-OC43, and feline coronavirus genomes were also cloned (Fig. 1). A total of 82 expression constructs, about one-third (25) of which originate from SARS-CoV and the rest from the other coronaviruses, were purified from yeast cells by using their GST tags. Immunoblot analysis revealed that most purified proteins could be detected and migrated at their expected molecular weights, including the glycoproteins.

To test whether a protein microarray approach could be used to

Conflict of interest statement: M.S. consults for Invitrogen.

Abbreviations: SARS, severe acute respiratory syndrome; CoV, coronavirus; IIFT, indirect immunofluorescence test; BCoV, bovine coronavirus; HCoV, human coronavirus; *k*-NN, *k* nearest neighbor; N, nucleocapsid; LR, logistics regression.

[†]H.Z., S.H., G.J., and X.Z. contributed equally to this work.

[§]Present address: Department of Pharmacology and Molecular Sciences and the HiT Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205.

^{††}To whom correspondence should be addressed. E-mail: michael.snyder@yale.edu.

© 2006 by The National Academy of Sciences of the USA

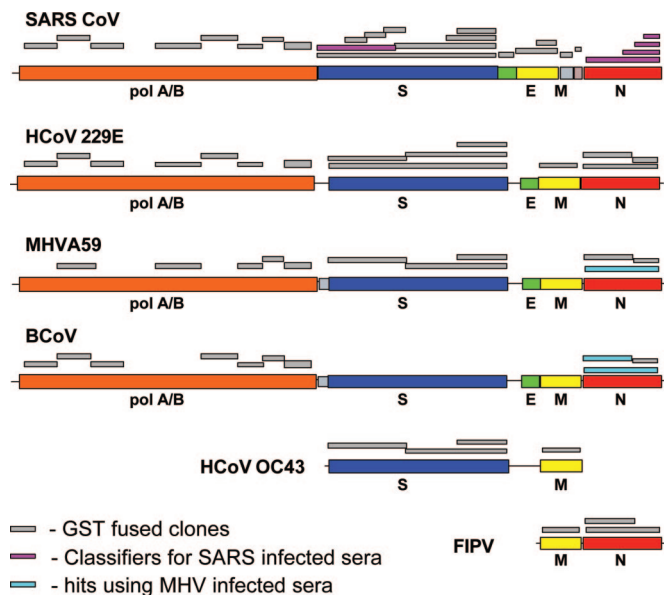


Fig. 1. Regions of six coronaviruses represented on the microarray. The positions of the cloned and expressed fragments are marked with light-gray bars. The pink bars represent SARS features selected as classifiers in the supervised cluster analysis (both *k*-NN and LR). The light-blue bars are features bound by the MHVA59-infected mouse serum.

detect SARS-CoV infection, we fabricated a microarray containing the 82 purified proteins. Serial dilutions prepared from four serum samples collected from Chinese patients clinically diagnosed as SARS-positive and which also tested positive by a local ELISA were used to probe the array. The presence of human-anti-SARS antibodies was detected with Cy3-labeled goat anti-human IgG antibodies (12–16). As shown in Fig. 2*A*, the sensitivity of the microarray assay is extremely high; reactivity is readily detected at 1:10,000-fold dilution for the strong positive serum and 1:800-fold for the weakly positive sera. The assay is 50-fold more sensitive than ELISAs performed using the same sera. Importantly, $<1 \mu\text{l}$ of serum is needed for the protein microarray assay, which is crucial because the sera from SARS patients are extremely precious.

Serum Probing of the Coronavirus Proteome Microarray with Human Sera. The coronavirus protein microarrays were used to screen sera from 399 Canadian and 203 Chinese infected and noninfected individuals in a double-blind format. The Canadian samples included 181 clinical- and laboratory-confirmed SARS-CoV sera (see *Materials and Methods*) (3), as well as anonymized clinical samples from patients who had presented with respiratory illness during the outbreak period but who failed to meet the case definition and did not develop SARS. Other SARS-CoV-negative sera were from asymptomatic healthcare workers. The Chinese sera were from patients with fever during the SARS outbreak; some of these were classified as SARS-positive and others, SARS-negative.

To accomplish the screening, each of the 82 purified coronavirus proteins was spotted in duplicate on eight identical blocks per microscope slide. Human IgG protein was also included as positive control (see below). The amount of immobilized coronavirus proteins and protein fragments present on the microarray was quantified by probing with anti-GST antibodies (Fig. 2*B*).

The serum samples were screened at a 200-fold dilution, and bound antibodies were detected with Cy3-labeled goat anti-human IgG. The signals were analyzed by using algorithms that we developed. Positive sera usually exhibited strong reactivity for $\approx 10\%$ of the proteins on the microarrays. The full-length and two C-terminal derivatives of SARS N-protein were strongly recognized by the

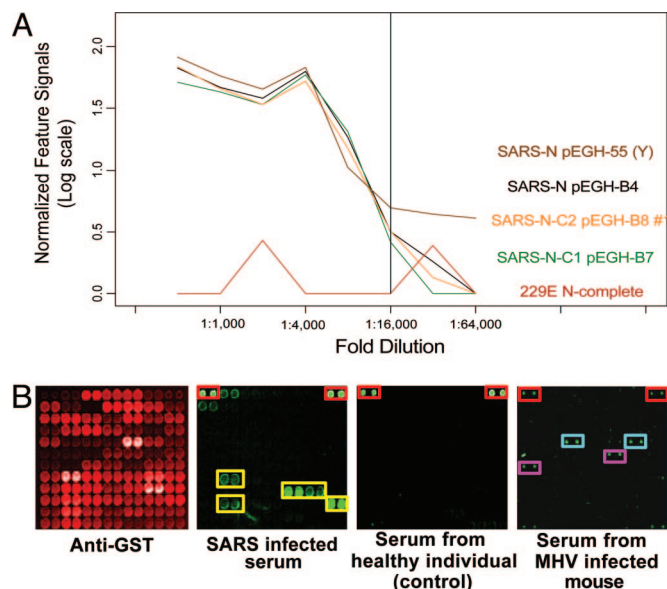


Fig. 2. Analysis of patient serum samples in a protein microarray format. (A) A SARS-CoV-positive serum from a diagnosed SARS-CoV-infected patient in Beijing was tested at eight dilutions. The signals for the five SARS N protein fragments are shown on the chart. The vertical line indicates the detection limit. (B) Examples of coronavirus protein microarrays probed with various sera from SARS-CoV-infected or uninfected individuals. The first image shows probing with an anti-GST antibody. The second image shows probing with a serum from a SARS patient. The N protein and its fragments were the most antigenic protein on the array [indicated by the yellow boxes (second image)]. The third image shows probing with a serum from a non-SARS patient. The fourth image shows probing with a serum from MHVA59-infected mouse. Light-blue boxes, the MHV N protein; pink boxes, the BCoV N protein. The red boxes indicate the signals from the human IgG used as the positive controls.

antibodies present in the SARS-CoV-infected patient sera but not in sera from noninfected individuals (Fig. 2*B*). The C-terminal fragments of the SARS N protein, which contains a short lysine-rich region (KTFPPTEKKDKKKKTDEAQ; amino acids 362–381) unique to SARS CoV, exhibit the highest antigenic activity (SARS-N-C2; Fig. 2*A* Right). These results are consistent with previous studies that identified the N proteins of coronaviruses as the most abundant and reactive antigens (11).

Although the N proteins are conserved among coronaviruses, the SARS-CoV-infected sera from the Chinese and Canadian patients showed little crossreactivity with proteins of other coronaviruses on the array, including N proteins. One exception is that many (88%) of the sera from the Chinese patients showed a slight reactivity to the first half of BCoV N protein, which shares $\approx 40\%$ identity through its first 210 amino acids with the SARS-CoV N protein. Interestingly, sera from infected Canadian patients did not react with this protein. In addition, $\approx 20\%$ of the sera from both SARS-positive and -negative Canadian individuals specifically recognized the HCoV-229E N protein but not the N proteins from the other species. We expect that many Canadian patients may have been exposed to HCoV-229E (see below).

To further test the specificity of our assays, we probed the coronavirus protein microarray with ≈ 30 sera from MHVA59-infected and control mice. As shown in Fig. 2*B*, a mouse-infected serum recognized the MHVA59 N protein, whereas control mouse sera did not react with proteins on the array. This serum also crossreacted with the N protein from BCoV and not with proteins from other coronaviruses. Because the N proteins from MHVA59 and BCoV share 70.7% identity and 87.9% similarity over their entire protein sequences, crossreactivity between these two proteins is not surprising.

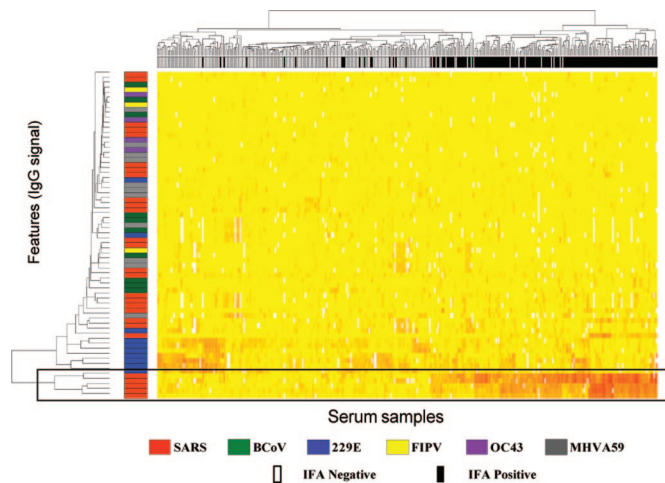


Fig. 3. Unsupervised 2D clustering of the Toronto sera and microarray features. The 399 Toronto IgG sera were clustered according to their reactivity to the microarray signals, and the microarray features were clustered according to their serum reactivity. The corresponding Euroimmun IIFT SARS-CoV IgG results are indicated on top of the diagram, where black and white bars represent SARS-positive and -negative sera, respectively. The different coronaviruses are color-coded on the left of the diagram. The yellow color is low or background signal on the arrays, whereas the orange color represents signals above the background level. The black box highlights the features that help classify SARS-infected sera from the microarray assays. All of the classifiers in the black rectangle are SARS N proteins and SARS N fragments.

In summary, although a few instances of crossreactivity occurred among highly similar proteins, the protein microarray approach demonstrated that different serum samples could be differentiated at a high degree of specificity. Most importantly, the protein microarray was able to distinguish reactivity between the human coronaviruses (HCoV-229E and SARS).

Detection of SARS-Infected Patients in the Canadian Samples. To determine whether an accurate SARS diagnostic test can be devised by using the protein microarray data, we analyzed the results obtained from the Canadian patients using computational approaches. The sera were first clustered according to the relative signal intensities of all of the coronavirus proteins immobilized on the microarrays in an unsupervised fashion (17). The sera fell into two major groups, which upon subsequent comparison with clinical IIFT data were largely correlated with SARS-positive and -negative sera (Fig. 3). The unsupervised method correctly predicted 138 of 181 infected serum samples (76% sensitivity, with sensitivity defined as the percentage of correct positives of the total positives) and 210 of 218 sera from healthy individuals (96% specificity, with specificity defined as the percentage of correctly classified negatives of the total negatives). In the cluster of markers, five of the SARS N protein fragments associated tightly (Fig. 3, at the bottom). Most of the sera clustered as originating from SARS-infected patients exhibited unambiguous reactivity with this group of markers as expected (Fig. 2B). The SARS sera also exhibited statistically significant binding to one spike protein fragment.

We next set out to improve our prediction by identifying the

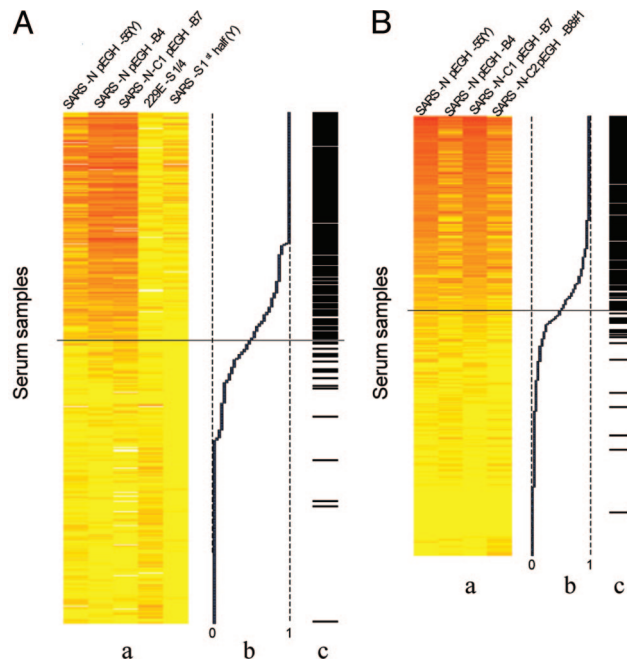


Fig. 4. Models generated by *k*-NN (A) and LR (B). The cutoff for the prediction is the probability of 0.5, which is indicated by the black horizontal line: (lane a) signals for the selected classifiers, (lane b) confidence calculated from the classifier signals (range from 0 to 1), and (lane c) the IIFT annotations, where the black and white boxes represent IIFT-positive and -negative, respectively. On the top are depicted the names of the features that were selected by the *k*-NN and LR models.

meaningful classifiers and conducting a supervised classification. Because only a limited number of proteins showed differences between the SARS-CoV-positive and -negative patients (Fig. 3), we selected the top 10 features that demonstrated the most significant differences between these two types of patients as candidates for classifier selection (18). Many of the selected candidates were SARS N protein fragments.

To determine the best classifiers and classification model, we applied two different supervised analysis approaches, *k* nearest neighbor (*k*-NN) (19) and logistics regression (LR) (20). *k*-NN measures the similarity between a new case and all of the known cases to make a prediction and is determined by the identities of its *k* closest neighbors (Fig. 4A). Using this method, five features were selected by the algorithm as the best classifiers: SARS N [pEGH-55 (Y)], SARS N (pEGH-B4), SARS N-C1 (pEGH-B7), 229E-S 1/4, and SARS spike [first half (Y)] (note that 229E-S1/4 negatively correlates with SARS). The best *k* value selected by the model is 9, indicating that the nine closest-neighboring samples to the tested case were used for the prediction. At the confidence cutoff of 0.5, this model achieved 91% accuracy with 15 positive and 18 negative cases missed [163 of 181 positive cases were correct (90% sensitivity) and 203 of 218 negative sera correct (93% specificity)] (Table 1).

We also analyzed our microarray results using LR, which is a generalized linear regression for binary responses (Fig. 4B). The features selected by LR included SARS N-C1 (pEGH-B7),

Table 1. Prediction performance of the two classification methods

	Number of cases	Correctly classified	False positive	False negative	Sensitivity, %	Specificity, %	Accuracy, %
<i>k</i> -NN	399	366	15	18	90	93	91
LR	371	359	12	18	89	94	92

SARS N (pEGH-55) (Y), SARS N (pEGH-B4), and SARS N-C2 (pEGH-B8 #1). The accuracy of this model was 92% (89% sensitivity and 94% specificity). To determine whether *k*-NN or LR performed better, we used the receiver operating characteristic curve (21) and plotted the rate of true positives against that of false positives at different cutoff points. Using the area under the curve (AUC), we measured the quality of the model and found that both AUC values were close to 0.95, indicating that both models performed equally well. Interestingly, although both LR and *k*-NN predictions exhibited only $\approx 92\%$ overlap with the IIFT results (Table 1), 97% of their predictions were shared, indicating that the discrepancy between our models and the standard IIFT test does not depend on the analysis method but rather on the experimental data.

That both *k*-NN and LR performed similarly prompted us to repeat the proings of the 33 discrepant sera along with some of those that agreed with the predictions. After these proings, eight reproducibly false-negative samples remained by both methods even after a third round of proings.

To test whether IgM would yield better results than IgG, particularly for patients during the acute phase of the disease, $\approx 90\%$ of the Toronto sera were also probed for IgM reactivity on the microarray. Except for one serum, the proings performed equal to or worse than the IgG proings, consistent with previous results (22–24).

Validation of the SARS Proteome Array Classification Method. To further examine the accuracy, sensitivity, and specificity of our approach, we conducted another double-blind experiment using 56 sera collected from Chinese patients; 36 of the patients were diagnosed as SARS-infected, and 20 were diagnosed as uninfected. All of the sera were collected from SARS patients who recovered from respiratory disease. Of the 56 serum samples, only one serum was misclassified by our models (98% accuracy, 100% sensitivity, and 95% specificity). Importantly, both the *k*-NN and the LR models predicted this serum to be positive with a confidence value of 1 on a 0 to 1 scale. Taken together, these results demonstrated that our prediction algorithms performed well and accurately identified the SARS-infected samples from a large population.

Comparing the Protein Microarray Results with ELISAs. To determine how the viral protein microarray compared with the current methods of diagnosis, we compared the performance of two independent ELISA tests on serum samples from both Canada and China. The Euroimmun ELISA was used on all but three of the serum samples taken from Canadian patients and resulted in two false-positive, six false-negative, and 26 borderline (uncertain/inconsistent) classifications. Thus, the Euroimmun ELISA is 91% accurate, as compared with 92% accuracy for the proteome array method. The samples missed by the two assays were not identical.

We also compared the microarray approach with a local ELISA used in China that used only the purified N protein. A set of 147 serum samples collected from fever patients during the SARS outbreak in China was used to probe the coronavirus protein microarray. The SARS status of these patients is not known. Similar to the results presented above, we found 85% agreement between the predictions made from the microarray assay and those made from the ELISA; all 70 sera that were SARS-CoV-positive by the ELISA were also positive by microarray. The microarray identified an additional 21 sera as SARS-CoV-positive that were not found by using the ELISA. Because (i) 15 of the 21 serum samples had confidence scores >0.72 , the lowest-confidence score for the 56 known Chinese SARS-infected sera presented above, and (ii) the rate of false positives in our assays is $<7\%$ (the overall specificity for the sera from characterized patients is $>99.56\%$), it is likely that most of

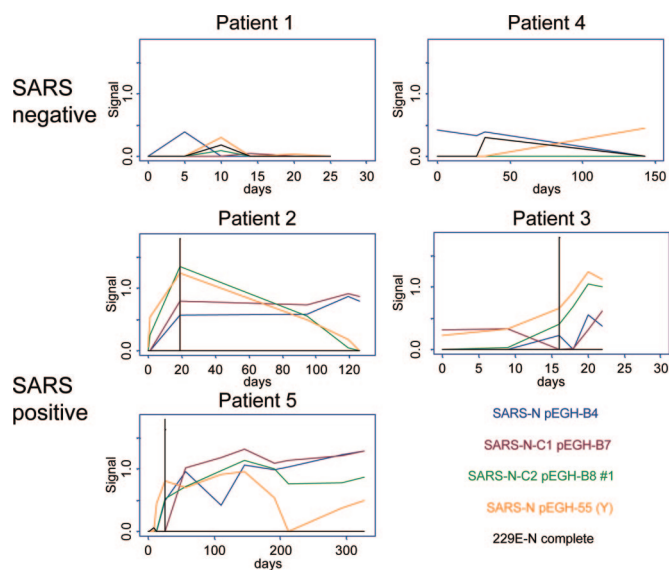


Fig. 5. Time-course analysis of serum reactivity of five Canadian individuals. (Top) Graphs from two individuals with non-SARS respiratory disease; (Bottom) Results from three SARS patients. The relative levels of antibodies against four of the SARS N protein constructs along with that of HCoV-229E N protein were monitored at different times. The vertical lines indicate the time at which the individuals were diagnosed as SARS-positive by biochemical assays.

these samples originated from SARS patients. In summary, these results indicate that the protein microarray method is at least as sensitive as the Euroimmun ELISA and more sensitive than the local Chinese ELISA, and therefore is an excellent assay for detecting SARS.

Anti-SARS Antibodies Can Persist Long After Initial Infection. One useful feature of a serum test relative to a nucleic acid diagnostic test is that anti-SARS antibodies can potentially be detected after infection. We therefore tested how long anti-SARS antibodies remained present in recovering patients after infection. Serum samples drawn from five Canadian individuals (two respiratory illness other than SARS and three confirmed SARS-CoV cases) at different times postinfection were tested by using the protein microarrays (Fig. 5). Reactivity to five N proteins (four SARS N proteins and one CoV-229E N protein) was scored. Sera from non-SARS patients (Patients 1 and 4 in Fig. 5) did not exhibit significant reactivity to any of the five SARS-CoV markers. In contrast, sera from SARS-CoV-positive patients (Patients 2, 3, and 5, Fig. 5) reacted strongly with each of the SARS N peptides, and for the two cases that were monitored over a long period (120–320 days), reactivity remained high for two N peptides. Furthermore, the above two SARS CoV N antigens were the same ones that reacted most strongly in the 36 SARS-confirmed patients from the group of 56 Chinese respiratory patients. These results demonstrate that at least some patients retain reactive antibodies for extended periods, and they can be detected by protein microarrays.

Extending the Protein Microarray Approach to Detecting Other Coronaviruses. Although this study was aimed at developing a systematic screen for SARS-infected sera, proteins from other human coronaviruses such as the HCoV-229E were included on the microarray, thus allowing the detection of antibodies directed toward other coronaviruses (25–27). Using 10 HCoV-229E-related proteins as classifiers, we identified 82 serum samples with substantial signal (52 of 218 SARS-CoV-negative (23.9%) and 30 of the 218 SARS-CoV-positive sera (13.8%). The presence of 52 HCoV-229E-positive sera in SARS-CoV-negative

patients suggests that these patients were or had been infected with HCoV-229E. The observation that many (150) patients are SARS-CoV-positive and lack HCoV-229E antibodies indicates that HCoV-229E and SARS-CoV infections can occur independently of each other. Because these sera were not tested for HCoV-229E infection, the number of false positives and negatives could not be scored. Nonetheless, these results indicate that our approach can likely be used to diagnose infections from related human coronaviruses.

Discussion

In this study, we present the construction and use of a coronavirus protein microarray to screen human sera for antibodies against human SARS and related coronaviruses. We tested >600 sera from two different parts of the world and predicted the nature of serum samples with >90% accuracy. To our knowledge, it is the largest study of this type conducted thus far, and the first to analyze patients from the two major geographical locations of the SARS epidemic. We compared our results with the current available methods and showed that the coronavirus protein microarray is at least as sensitive as and more specific than the available ELISA tests and has the advantage that multiple antigens from different coronavirus are tested simultaneously. Thus, this system has enormous potential to be used as an epidemiological tool to screen human and other sera for many types of viral infections as well as other types of disease (e.g., cancer).

Sensitivity and Accuracy of the Protein Microarray Assay. Using the Euroimmun IIFT plus epidemiological data as reference, the protein microarray assay offered several advantages relative to the commercially available Euroimmun ELISA. First, the assays were sensitive and functioned at high dilutions, allowing small amounts of sera to be used (1/200 dilution was used here instead of the 1/50 commonly used in ELISAs). This is particularly important for SARS research, because the sera are extremely precious and not replaceable. Consistent with an increased sensitivity, more Chinese patients were diagnosed as SARS-positive by using the protein microarray over the Chinese ELISA. Second, the accuracy of our assay is as good as, if not better than, the Euroimmun ELISA: 92% vs. 91% accuracy. Third, our assay has greater reliability, in that multiple antigens are followed, and a weighted scoring scheme based on probabilities was developed, instead of relying on the results of one or a mix of antigens. To our knowledge, a probabilistic test of this type has not been described previously for viral detection using sera, and we expect this approach to be of general utility. Fourth, our assay can monitor the presence of antibodies to multiple viruses allowing their potential simultaneous detection. Fifth, our assay can be automated to robotically probe hundreds of sera in parallel, a major advantage over the visual analysis in IIFT. Finally, unlike IIFT, in which results can be masked by the presence of high concentrations of antinuclear factor (60 such patients were present in our study), the protein array is not affected by such antibodies.

One concern with using protein microarrays is the reproducibility of the assay. After unblinding of the initial screening, we retested the ≈ 30 sera that exhibited either false-positive or -negative reactions; 22 were correctly reclassified. Furthermore, retesting 97 sera that were correctly classified but were close to the borderline resulted in misclassification of 13%. These results indicate that the assay as performed is 90% reproducible. The reason for this variation is currently unclear. Probing sera in triplicate will increase the reproducibility of the assay to 98% if the majority results are scored.

A subset of eight sera yielded false-negative results, whereas the patients had been classified as SARS-CoV cases using clinical and laboratory tests. This misclassification by the protein

microarray assay occurred regardless of the array interpretation method used. We presume that either these patients were misclassified clinically, or IIFT is a more sensitive assay than the protein microarray. Possible explanations for the latter include that IIFT was tested at a lower serum dilution (1/10) as compared to the arrays (1/200), or that the SARS proteins had been purified from yeast cells, which have different posttranslational modifications compared with those of mammalian cells. Some sera may recognize glycosylated antigens modified in humans that are not present on the antigens prepared in yeast (see ref. 28). Consistent with this hypothesis, the infected sera primarily recognized the SARS-CoV-encapsulated N protein but none of the six surface glycoproteins. The purification of viral proteins from human cell lines should relieve this problem.

Specificity of the Coronavirus Microarray for Detecting Different Viral Infections. Most of the human sera did not crossreact with antigens from other species, indicating the assay is specific. However, 82 individuals had antibodies reactive to HCoV-229E antigens. These were observed both in SARS-CoV-positive and -negative patients. Because these antibodies were observed in both types of patients, the simplest explanation is that these patients were exposed to HCoV-229E (or a closely related virus). It is unlikely that the antibodies present in SARS-CoV-infected patients crossreact with HCoV-229E antigens, because HCoV-229E and SARS-CoV belong to different phylogenetic groups, and their N antigens are only 27% identical. Thus, we expect our protein microarray assay monitors exposure to several types of coronaviruses.

In summary, we have constructed coronavirus protein microarrays that cover proteins from six coronavirus proteomes and have used them to classify sera from potential SARS-infected patients. The approaches developed here are applicable to potentially all viruses and are expected to have great impact in epidemiological studies and possibly in clinical diagnosis.

Materials and Methods

Serum Samples. The 399 serum samples tested from Canada included 40 acute and 164 convalescent sera from 92 patients who met the clinical and laboratory criteria for SARS-CoV infection during the 2003 Toronto SARS outbreak. Sera from 112 Toronto patients who presented with non-SARS respiratory illness and 83 sera from health professionals were also included. None of the acute, all 164 of the convalescent, and 17 of the sera from 12 healthcare workers demonstrated IgG antibodies as detected by using the Euroimmun IIFT test. All positive results were repeated, and any unexpected result was confirmed by using the SARS-CoV neutralization assay. The Chinese samples were collected from several hospitals in Beijing by the Beijing Genomics Institute. These sera were collected from 147 non-confirmed fever and 56 respiratory patients (36 confirmed SARS patients and 20 non-SARS individuals).

Preparation of a Coronavirus Microarray. The SARS ORFs were amplified by RT-PCR from the SARS-CoV isolate BJ01 (GenBank accession no. AY278488) and cloned into a yeast GST expression vector (pEGH) described previously (12). The same approach was used for the cloning of other coronavirus genes. All clones were confirmed by sequencing their inserts.

The constructs were transformed into yeast, and proteins were purified as described (13). For samples that exhibited low yields, the purification was repeated by using 50-ml cultures and/or up to four purifications. The coronavirus protein microarrays were fabricated by spotting the purified proteins along with positive control proteins onto eight-pad FAST slides (Schleicher & Schuell) using a microarrayer (Bio-Rad). The printed arrays were incubated overnight at 4°C and stored at -20°C.

Serum Assays on Coronavirus Protein Microarrays. An eight-hole rubber gasket (Schleicher & Schuell) was applied to each microarray to form eight individual chambers. The surfaces were blocked (SuperBlock; Pierce) at room temperature (RT). Each serum sample was diluted 200-fold in SuperBlock and incubated on microarrays at RT for 1 hour with gentle shaking. The Chinese sera were further filtered before probing the arrays. After gasket removal, the microarrays were washed extensively in a large volume of PBS wash buffer with shaking. To visualize the presence of human antibodies, Cy3- and Cy5-labeled anti-human IgG and IgM antibodies (The Jackson Laboratory) were incubated at 1,000-fold dilution. The arrays were washed with PBS buffer, briefly rinsed with water, and dried. The slides were scanned and signals analyzed using GENEPIX PRO 3.0 software (Molecular Devices).

Reproducibility of the assay was examined both in a blinded and an unblinded fashion. First, multiple aliquots of 14 sera from 13 patients were embedded into the serum selection for a total of 32 samples; 19 and 13 sera were derived from SARS-CoV-positive and non-SARS individuals, respectively. Each sample was repeated at least once. Upon unblinding, the IIFT results were compared with those obtained by the arrays. Second, the results obtained from 111 convalescent sera drawn on different dates from 35 SARS-positive patients (2–11 specimens per patient) and for 32 convalescent sera received from seven non-SARS individuals (two to nine specimens) were evaluated by comparison with those from the microarray probing assays. Array results correlated within patients and agreed for all 70 sera received from 23 of 35 SARS-CoV-positive patients, one of whom had a series of 11 positive samples from different dates over nearly 1 year of followup. However, for 6 of 35 patient series (20 samples), a single sample per patient yielded a discrepant negative result by arrays and in a further five patient series (15 samples), two samples gave false negatives. For the unblinded method, one of four of the serum samples (97) that were classified correctly and near the borderline were probed a second time. Approximately 90% yielded results similar to the first probings.

Data Normalization and Hierarchical Clustering. Given the nature of each serum we collected, we expected a wide range of antibody titer. To compensate for this effect in the final clustering and classification, we log-transformed the intensities and then normalized the numbers in a way that each probing had the same median and median absolute deviation values (see supporting information, which is published on the PNAS web site). Divisive hierarchical clustering was then applied to both the sera and the array features by using S-PLUS 6.1 (17).

k-NN. *k*-NN stores a group of known cases and classifies new instances based on a similarity measure (19). The new instance is classified according to the identities of its nearest neighbors. The number of neighbors is determined by the parameter *k*, and the similarity is measured as the Euclidean distance by using the signals of the classifiers. The best parameters were selected in the learning process and applied in the predicting process. In the learning process, all parameters, including possible *ks*, and candidate classifiers were tested and their performance evaluated at 10-fold crossvalidation to find the best values (29). In the prediction process, the *k*-NNs were retrieved for each new instance, and classifications were made according to the memberships of the neighbors.

LR. LR is a generalized linear regression model designed for binary responses (20). However, no missing values for the candidate features are allowed in model construction; thus, the number of sera analyzed (≈ 370) was less than the total screened.

The candidate features were selected by the model using both direction stepwise search with Akaike information criterion (30). We performed this analysis using S-PLUS 6.1 software that selected the top four features out of the candidate list for the prediction step. Finally, the probability of each serum to be positive was calculated by using those features, and those that had a value >0.5 were classified as SARS-CoV-positive.

We thank J. Mok and K. Wise for critical reading of the manuscript. This research was supported by a National Institutes of Health grant. H.Z. was supported by a Damon Runyon Foundation postdoctoral fellowship.

- Donnelly, C. A., Ghani, A. C., Leung, G. M., Hedley, A. J., Fraser, C., Riley, S., Abu-Raddad, L. J., Ho, L. M., Thach, T. Q., Chau, P., et al. (2003) *Lancet* **361**, 1761–1766.
- McIntosh, K., Chao, R. K., Krause, H. E., Wasil, R., Mocega, H. E. & Mufson, M. A. (1974) *J. Infect. Dis.* **130**, 502–507.
- Rainer, T. H., Cameron, P. A., Smit, D., Ong, K. L., Hung, A. N. W., Nin, D. C. P., Ahuja, A. T., Si, L. C. Y. & Sung, J. J. Y. (2003) *Br. Med. J.* **326**, 1354–1358.
- Bressler, A. M. & Nolte, F. S. (2004) *J. Clin. Microbiol.* **42**, 987–991.
- Peiris, J. S., Chu, C. M., Cheng, V. C., Chan, K. S., Hung, I. F., Poon, L. L., Law, K. I., Tang, B. S., Hon, T. Y., Chan, C. S., et al. (2003) *Lancet* **361**, 1767–1772.
- Peiris, J. S., Lai, S. T., Poon, L. L., Guan, Y., Yam, L. Y., Lim, W., Nicholls, J., Yee, W. K., Yan, W. W., Cheung, M. T., et al. (2003) *Lancet* **361**, 1319–1325.
- Poon, L. L., Chan, K. H., Wong, O. K., Yam, W. C., Yuen, K. Y., Guan, Y., Lo, Y. M. & Peiris, J. S. (2003) *J. Clin. Virol.* **28**, 233–238.
- Wang, J., Wen, J., Li, J., Yin, J., Zhu, Q., Wang, H., Yang, Y., Qin, E., You, B., Li, W., et al. (2003) *Clin. Chem.* **49**, 1989–1996.
- Shi, Y., Yi, Y., Li, P., Kuang, T., Li, L., Dong, M., Ma, Q. & Cao, C. (2003) *J. Clin. Microbiol.* **41**, 5781–5782.
- Guan, M., Chan, K. H., Peiris, J. S., Kwan, S. W., Lam, S. Y., Pang, C. M., Chu, K. W., Chan, K. M., Chen, H. Y., Phuach, E. B., et al. (2004) *Clin. Diag. Lab. Immunol.* **11**, 699–703.
- Leung, D. T., Tam, F. C., Ma, C. H., Chan, P. K., Cheung, J. L., Niu, H., Tam, J. S. & Lim, P. L. (2004) *J. Infect. Dis.* **190**, 379–386.
- Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., et al. (2001) *Science* **293**, 2101–2105.
- Zhu, H., Klemic, J. F., Chang, S., Bertone, P., Casamayor, A., Klemic, K. G., Smith, D., Gerstein, M., Reed, M. A. & Snyder, M. (2000) *Nat. Genet.* **26**, 283–289.
- MacBeath, G. & Schreiber, S. L. (2000) *Science* **289**, 1760–1763.
- Haab, B. B., Dunham, M. J. & Brown, P. O. (2001) *Genome Biol.* **2**, research0004.
- Joos, T. O., Schrenk, M., Hopfl, P., Kroger, K., Chowdhury, U., Stoll, D., Schorner, D., Durr, M., Herick, K., Rupp, S., et al. (2000) *Electrophoresis* **21**, 2641–2650.
- Struyf, A., Hubert, M. & Rousseeuw, P. J. (1997) *Comput. Stat. Data Anal.* **26**, 17–37.
- Dudoit, S., Fridlyand, J. & Speed, T. P. (2002) *J. Am. Stat. Assoc.* **97**, 77–87.
- Cover, T. (1967) *IEEE Trans. Inf. Theor.* **13**, 21–27.
- Stynes, D. & Peterson, G. (1984) *J. Leisure Res.* **16**, 295–310.
- Goddard, M. J. & Hinberg, I. (1990) *Stat. Med.* **9**, 325–337.
- Liu, I. J., Hsueh, P. R., Lin, C. T., Chiu, C. Y., Kao, C. L., Liao, M. Y. & Wu, H. C. (2004) *J. Infect. Dis.* **190**, 797–809.
- Hsueh, P. R., Huang, L. M., Chen, P. J., Kao, C. L. & Yang, P. C. (2004) *Clin. Microbiol. Infect.* **10**, 1062–1066.
- Woo, P. C., Lau, S. K., Wong, B. H., Chan, K. H., Chu, C. M., Tsoi, H. W., Huang, Y., Peiris, J. S. & Yuen, K. Y. (2004) *Clin. Diag. Lab. Immunol.* **11**, 665–668.
- Macnaughton, M. R., Hasony, H. J., Madge, M. H. & Reed, S. E. (1981) *Infect. Immun.* **31**, 845–849.
- Cereda, P. M., Pagani, L. & Romero, E. (1986) *Eur. J. Epidemiol.* **2**, 112–117.
- Pohl-Koppe, A., Raabe, T., Siddell, S. G. & ter Meulen, V. (1995) *J. Virol. Methods* **55**, 175–183.
- Wildt, S. & Gerngross, T. U. (2005) *Nat. Rev. Microbiol.* **3**, 119–128.
- Mullin, M. & Sukthankar, R. (2000) in *Proc. Int. Conf. Machine Learn.*, 639–646.
- Akaike, H. (1974) *IEEE Trans. Aut. Control* **AC19**, 716–723.