An Efficient Primary Screening of COVID-19 by Serum Raman Spectroscopy

Gang Yin,1,2#, Lintao Li, 1,2#, Shun Lu1,2#, Yu Yin3 ,Yuanzhang Su4, Yilan Zeng5, Mei Luo5, Maohua Ma5, Hongyan Zhou1,2,Lucia.orlandini1,2，Dezhong Yao3\*, Gang Liu6\*, Jinyi Lang1,2\*.

1Department of radiation oncology, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, Chengdu , China.

2Radiation Oncology Key Laboratory of Sichuan Province, Chengdu, China.

3Sichuan Institute for Brain Science and Brain-Inspired Intelligence，MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, Chengdu, China.

4School of Foreign Languages, University of Electronic Science and Technology of China, Chengdu, China.

5The Public Health Clinical Center of Chengdu, Chengdu, China.

6The first affiliated hospital of Chengdu medical college, Chengdu, China.

# These three authors contribute equal to this manuscript

\* Corresponding author

Jinyi Lang, Dezhong Yao and Gang Liu are all the corresponding author of the study.

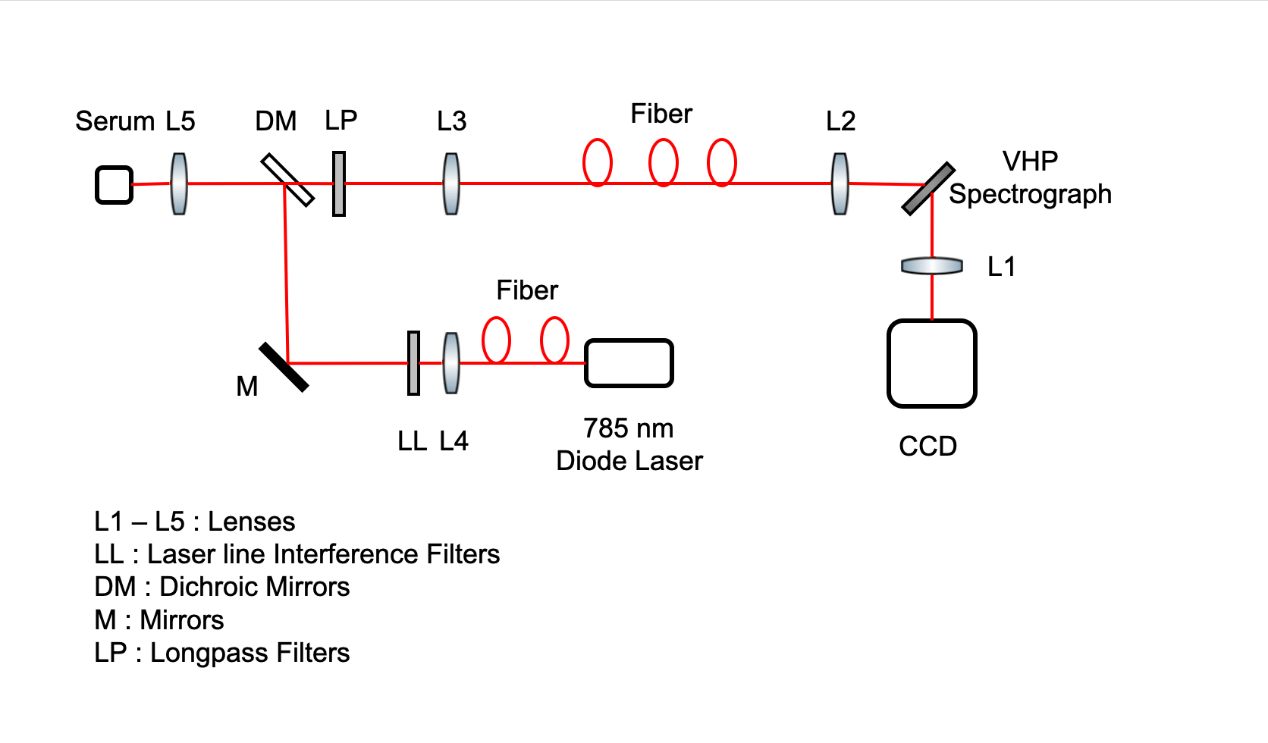
To whom correspondence should be addressed. Prof Jinyi Lang, Tel: +8602885420241; E-mail: [langjy610@sichuancancer.org](mailto:langjy610@sichuancancer.org)

**supplementary material**

1. **supplementary material 1**

The Raman system consists of a Volume Phase Holographic (VPH) spectrograph, a deep-cooled camera (CCD, a Raman probe, and a laser (Supplementary Figure1: detailed optical configuration and physical scheme). It is designed by Sichuan Institute for Brain Science and Brain-Inspired Intelligence, Chengdu, 611731, China.

The Raman probe has three ENDS: sample END, laser END, and spectrograph END. The laser END has a SMA fiber adaptor and collimating lens to get collimated 785 nm laser. An internal laser line filter (Semrock, LL01-785-12·5) was applied to obtain a clean laser profile. The sample end was customized to hold a blood serum container using a microscope objective (50X, NA 0·5, WD 8·0, Sunnyoptical) as a focusing lens. The spectrograph end has SMA fiber adaptor and sequential dual-edge filters designed to block all back-scattered Reighley signal from the sample, allowing only Raman signal to pass. A lens was used to focus the Raman signal into an optical fiber with a core diameter of 300 um . The fiber directs the signal into a spectrograph (lens based VPH grating type, F/2·2, EMvision） with a thermoelectric-cooled CCD camera (at -60 °C, Andor iVac DR-316B-LDC-DD). A single-mode diode laser (real-light) with wavelength 785 nm and power 100 mW was used for Raman excitation. The laser power on the sample was detected around 70 mW. Spectra were recorded in the spectral range 600-1800 cm–1. The spectra was collected with approximately 15 scans per serum sample with a 3 s accumulation for each scan. Data contaminated by cosmic rays were removed from the spectra.



Supplementary Figure1. Detailed optical configuration and physical figure

1. **supplementary material 2**

**Experiment and Analysis.**

First, the spectrograph wavenumber calibration was performed with reference bands of toluene[1](#_ENREF_1); at the same time, ethanol spectrum was measured using an exposure time of 3 s for each wavenumber calibration experiment. Secondly, the spectra was acquired with cryopreservation tubes with 5% normal saline using an exposure time of 3 s and with five successive scans for every beginning and finish the experiment. The average spectra obtained with the cryopreservation tubes was used to investigate the influence of tube wall materials. Next, the Raman spectra of the serum samples sealed with cryopreservation tube were measured using the same integration parameters as with the cryopreservation tube measurements. Three experimenters took the Raman scan for each sample tube, and repeated five times the experiment. The cryopreservation tube was placed in a special cards slot of the Raman spectrometer to ensure that the laser path through the tube wall was at a certain angle.

**Data processing steps**

1）Smoothing; a Savitzky–Golay digital–moving average filter was applied over a span of seven points with a 4th order polynomial fitting[2](#_ENREF_2), [3](#_ENREF_3);

2）Baseline correction: the Improved Modified Multi-Polynomial Fitting algorithm was used[4](#_ENREF_4);

3）A normalization by total area was performed after subtracting the fluorescent background;

4）ANOVA analysis was used for the statistical test; each wavenumber data point was considered as a feature to be examined. For each compared group, wavenumber points of intra-group were analysed to verify if their distribution satisfied Gaussian distribution; the homogeneity of variances of inter-group was then evaluated. Random sampling was repeated one hundred timed, and for each random sampling 70% of the data selected randomly passed Gaussian distribution test; homogeneity of variances test was selected for ANOVA test (Statistical significance level p < 0·05). Points showing statistical significance of ANOVA test for more than 70 times out of 100, were selected as the features to input in machine learning.

5）The procedure described above (4) was used to process the data of inter-group validation.

**Support vector machine (SVM)**

Most of the time, different biomolecules will cause the same Raman peak. Assigning these peaks to any particular macromolecule can cause inaccurate results. The processed data is then used for classification purposes. In this case, machine learning algorithms are very effective for analyzing Raman spectrum data. SVM is considered as an effective classifier because it can not only classify the data, but also optimize the decision boundary by maximizing the margin between data clusters. SVM can be used for linearly separable and linearly inseparable data sets. It is a powerful discriminative learning algorithm that has aroused great interest in the field of biomedical optics[5](#_ENREF_5). SVM and Raman spectroscopy have previously been used to isolate dengue infected sera from normal healthy sera[3](#_ENREF_3). The combination of SVM and leave-one-out cross-validation chemometric methods for the classification of patients with liver cirrhosis has been previously reported[6](#_ENREF_6). In our SVM method, a non-linear radial basis function (RBF) is used.

In this study, the feature selection method uses the ANOVA statistical test method (explained in detail in the Data processing steps). In the classification model, 82 points, 50 points, and 14 points, passed the ANOVA test for suspected vs COVID-19, healthy controls vs COVID-19, and healthy controls vs suspected, respectively. The points that passed the ANOVA statistical test between the two groups were used as input features of the SVM.

**Model evaluation**

After ANOVA test, the study was carried a step further to examine the performance of the machine learning algorithms on the classification of the Raman spectral data. Based on the domain knowledge, the model predicts unknown samples and classifies them accordingly. The performance of the proposed model has been evaluated using a cross-validation method. The cross-validation method first divides the whole data set in two groups containing 70% and 30% of the whole data, for training and testing groups, respectively. The whole data set include measurements from 53, 54 confirmed and suspected, respectively, COVID-19 patients, and from 50 healthy controls individuals. The training and the cross-validation data sets were separated by selecting randomly 70% of the total data (measurements from 37 COVID-19 confirmed, 38 suspected, and 35 healthy indivuals). The remaining 30% of the data (measurements from 16 COVID-19, 19 suspected, and 15 healthy controls) were used as unseen data to assess the predictive power of the classification models; repeating 50 times the same process, we obtain a total of 50 models.

The models were built based on single spectrum (each serum sample, we get 15 Raman spectrum), as we intend to test if a single serum sample is affected by COVID-19 or not. Under this condition, 15 spectra of each serum sample were evaluated.

**Impendent dataset test**

To test the SVM model’s performance 50 SVM models based on the result of the training SVM model; of these 50 SVM models, 11 randomly selected were used to discriminate 15 spectra for each serum sample of 20 serum samples; the procedure was repeated 20 times 3300 judgments per serum sample, and the classification of each spectrum was recorded. All unlabeled spectra was assigned by SVM model to the class with the highest probability. True classification of the samples was only revealed after the model had made its predictions.

To check whether patients were affected by COVID-19 virus or not, two models were used: the COVID-19 vs. suspected model for people with symptoms, and the control model COVID-19 vs. healthy, for the asymptomatic patients.

References:

1. John R. Ferraro KNaCWB. Introductory Raman Spectroscopy. *Academic Press, Amsterdam, Second Edition,* 2003.

2. P L. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemometrics & Intelligent Laboratory Systems* 2012: 117(none):00-4.

3. Wang Q. Raman spectroscopic characterization and analysis of agricultural and biological systems. 2013.

4. Zhao J, Lui H, McLean DI, Zeng H. Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy. *Applied spectroscopy* 2007; **61**(11): 1225-32.

5. Sattlecker M, Bessant C, Smith J, Stone N. Investigation of support vector machines and Raman spectroscopy for lymph node diagnostics. *The Analyst* 2010; **135**(5): 895-901.

6. Zhang X, Thiefin G, Gobinet C, et al. Profiling serologic biomarkers in cirrhotic patients via high-throughput Fourier transform infrared spectroscopy: toward a new diagnostic tool of hepatocellular carcinoma. *Translational research : the journal of laboratory and clinical medicine* 2013; **162**(5): 279-86.