SUPPORTING INFORMATION

Integrative Modelling of Quantitative Plasma Lipoprotein Sub-class, Glycoprotein and Amino Acid Metabolic Data Reveals a Multisystem Signature of SARS-CoV-2 Infection

Torben Kimhofer[1], Samantha Lodge[1], Luke Whiley[1], Nicola Gray[1], Ruey Leng Loo[1], Nathan G Lawler[1], Philipp Nitschke[1], Sze-How Bong[1], David L. Morrison[1], Sofina Begum[2], Toby Richards[3], Bu B. Yeap[3], Chris Smith[4], Kenneth G. C. Smith[4], Elaine Holmes[1,2*] and Jeremy K. Nicholson[1,5*]

[1]Australian National Phenome Centre, Computational and Systems Medicine, Health Futures Institute, Murdoch University, Harry Perkins Building, Perth, Australia, WA6150.

[2]Department of Metabolism, Digestion and Reproduction, Imperial College London, Sir Alexander Fleming Building, South Kensington, London SW72AZ, UK.

[3]Medical School, Faculty of Health and Medical Sciences, University of Western Australia, and Department of Endocrinology and Diabetes, Fiona Stanley Hospital, Harry Perkins Building, Murdoch, Perth, Australia, WA6150.

[4]The Cambridge Institute of Therapeutic Immunology and Infectious Disease, Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge UK.

[5]Institute of Global Health Innovation, Imperial College London, Level 1, Faculty Building South Kensington Campus, LondonSW72NA UK

Correspondence to: Jeremy.nicholson@murdoch.edu.au, j.nicholson@imperial.ac.uk; elaine.holmes@murdoch.edu.au., elaine.holmes@imperial.ac.uk

Acknowledgements: We thank Spinnaker Foundation, WA, The McCusker Foundation, WA, The Western Australian State Government, and the MRFF for funding the Australian National Phenome Centre for this and related work. We thank the UK MRC for funding (SB), and the WA Premiers Fellowship funding for EH and RLL. We thank the Australian Research Council funding EH as an ARC Laureate Fellow.

1. Data modelling strategy

NMR and MS-derived data were combined and interrogated using principal components analysis (PCA) and orthogonal-partial least squares (O-PLS) as unsupervised and supervised multivariate analysis techniques, respectively. Data were mean-centred and auto-scaled prior to multivariate modelling.

OPLS-DA Model Training: The training sample set comprised a single time point from seven patients who tested SARS-CoV-19 RNA positive by a PCR swab test. Eight healthy controls were matched in sex and age, COVID-19 negativity was established serologically by double negative outcome in Anti-SARS-CoV-2 IgG and IgA ELISA from EUROIMMUN (Lübeck, Germany).

OPLS-DA model generalisability was estimated with the area under the receiver operator characteristic curve, determined using the predictive component scores projections from internal leave-one-out cross-validation ($AUROC_{CV}$). The $AUROC_{CV}$ amounted to a value of 1, indicating excellent model generalisability, and the amount of variation in metabolite data matrix explained by the model (R^2X) amounted to 25%.

External Model Validation: A separate sample set was used for model validation, comprising 11 patient samples (all tested SARS-CoV-19 RNA positive) and 17 healthy control participants. For the O-PLS-DA model with single column Y-block and 1 orthogonal + 1 predictive component, scores projections (*t'*) and outcome predictions ($\hat{y}_{OBJ}$) were calculated in the following multi-step process:

$$t'_o = (X'\, w_o) \,/\, |w_o|$$

with t'_o representing orthogonal component scores derived from auto-scaled validation data *X'* (using mean and standard deviation of the training data set) and O-PLS-DA model weights of the orthogonal component, w_0. Systematic Y-orthogonal variation is then removed from the validation data forming X'_r:

$$X'_r = X' - (t'_o\, p_o^T)$$

The filtered data are the basis for the calculation of new predictive component scores, t'_p using the O-PLS-DA model weights w_p:

$$t'_p = X'_r \, w_p$$

Finally, binary outcome model predictions ($\hat{y}$) are derived by weighting t'_p with the inner relation coefficient β and back-scaling using the training set-derived standard deviations (*s*) and means ($\bar{x}$):

$$\tilde{y} = \left[\left(\beta t'_p \right) s \right] + \bar{x}$$

The class prediction using the dummy variable-encoded outcomes (healthy = 0 and SARS-Cov-2 positive = 1) were performed as follows:

$$\hat{y} = \begin{cases} 0, & if \ \tilde{y} \leq 0.5 \\ 1, & otherwise \end{cases}$$

The model's prediction capacity was characterised by comparing the binary class predictions ($\hat{y}$) with the observed class outcomes (*y*) using accuracy, sensitivity, specificity and positive and negative prediction values (Section S1).

Variable Importance Calculation: The discriminatory power of variables was visualised by combining predictive component loadings (p_{pred}) from the OPLS-DA model, with p values derived from statistical group comparison using Kruskal Wallis Rank Sum test ($\alpha = 0.05$, two-tailed and FDR-adjusted), and Cliff's delta (Cd) as a non-parametric effect size measure. Benjamini-Yekutieli's method was chosen as multiple testing method as it accommodates dependencies among metabolic variables, e.g., biological dependencies among certain amino acids as well as plasma lipoprotein concentrations that were derived from the deconvolution of the same NMR signals. Cd takes values that range between -1 and 1, with low (high) magnitudes indicating a high (low) degree of overlap between distributions across two groups, and the sign indicating the direction of change conditioned to a reference group:

$$Cd = \frac{\#(x_i > x_j) - \#(x_i < x_j)}{mn}$$

With # denoting the number of times a value is higher in group A where $\forall i \in \{i, \dots, m\}$ or in group B where $\forall j \in \{1, \dots, n\}$. Throughout this work the healthy group was used as reference.

2. Indices used to assess prediction capacity of the O-PLS-DA model

The prediction capacity of the O-PLS-DA model was characterised by comparing the binary class predictions ($\hat{y}$) with the observed class outcomes (*y*) of the validation data set. Based on the characteristic contingency table (Table S1), the following indices were calculated:

$$Accuracy = \frac{TP + TN}{TOTAL}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Positive\ Predictive\ Value\ (PPN) = \frac{TP}{TP + FP}$$

$$Negative\ Predictive\ Value\ (NPN) = \frac{TN}{TN + FN}$$

With *TOTAL* representing the total number of validation data instances (*TOTAL=TN+FN+FP+TP*).

Table S1. Full cohort demographic data

	SARS-Cov2 (n=17)	Healthy Controls (n=25)
Male Sex	11 (68.75%)	17 [70.5%]
Age, yrs [SD]	69.4 [±10.5]	48.8 [± 15.3]
BMI, kg/m^2 [SD]	32.9 [±9.0]	25.9 [± 2.8]
Diabetes (T1/T2)	3 (12.5%)	0
Hypertension	9 [47%]	4 (16%)
Asthma	3 [17.6%]	0
COPD*	1 [5.9%]	0
Arthritis	4 [23.6%]	0
Glaucoma	2 [11.8%]	0
Dyslipidaemia	1 [5.9%]	1 (4%)
Chronic Renal Disease	1 [5.9%]	0
Chronic Heart Disease	4 [23.5%]	0

* Chronic obstructive pulmonary disease

Table S2: Symptoms presentation in SARS-CoV-2 positive patients

Symptom	SARS-Cov2 (n=17)
Fever	11 (64.7%)
Cough	9 (52.9%)
Shortness of Breath	6 (35.3%)
Sore Throat	1 (5.9%)
Rhinorrhoea	2 (11.8%)
Wheeze	1 (5.9%)
Chest Pain	1 (5.9%)
Myalgia	2 (11.8%)
Joint Pain	1 (5.9%)
Fatigue	5 (29.4%)
Headache	2 (11.8%)
Confusion	2 (11.8%)
Abdominal Pain	1 (5.9%)
Vomit	2 (11.8%)
Diarrhoea	4 (23.5)
Conjunctivitis	0.0
Lymphadenopathy	0.0

Table S3. Annotation of the keys used by the Bruker IVDr Lipoprotein Subclass Analysis (B.I.-LISA ™) method. Abbreviations: LDL – low-density lipoprotein; HDL – high-density lipoprotein; VLDL – very low-density lipoprotein; IDL – intermediate-density lipoprotein.

Key	Class / Subclass	Compound	Unit
TPTG	Total Plasma	Triglycerides	mg/dL
TPCH	Total Plasma	Cholesterol	mg/dL
LDCH	LDL	Cholesterol	mg/dL
HDCH	HDL	Cholesterol	mg/dL
TPA1	Total Plasma	Apolipoprotein-A1	mg/dL
TPA2	Total Plasma	Apolipoprotein-A2	mg/dL
TPAB	Total Plasma	Apolipoprotein-B100	mg/dL
LDHD	Ratio LDL and HDL Cholesterol	LDL Cholesterol / HDL Cholesterol	-/-
ABA1	Ratio of Apolipoproteins A1 and B100	Apolipoprotein-A1 / Apolipoprotein-B100	-/-
TBPN	Apolipoprotein-B100 carrying particles	Particle Number	nmol/L
VLPN	VLDL	Particle Number	nmol/L
IDPN	IDL	Particle Number	nmol/L
LDPN	LDL	Particle Number	nmol/L
L1PN	LDL-1	Particle Number	nmol/L
L2PN	LDL-2	Particle Number	nmol/L
L3PN	LDL-3	Particle Number	nmol/L
L4PN	LDL-4	Particle Number	nmol/L
L5PN	LDL-5	Particle Number	nmol/L
L6PN	LDL-6	Particle Number	nmol/L
VLTG	VLDL Class	Triglycerides	mg/dL
IDTG	IDL Class	Triglycerides	mg/dL
LDTG	LDL Class	Triglycerides	mg/dL
HDTG	HDL Class	Triglycerides	mg/dL
VLCH	VLDL Class	Cholesterol	mg/dL
IDCH	IDL Class	Cholesterol	mg/dL
LDCH	LDL Class	Cholesterol	mg/dL
HDCH	HDL Class	Cholesterol	mg/dL
VLFC	VLDL Class	Free Cholesterol	mg/dL
IDFC	IDL Class	Free Cholesterol	mg/dL
LDFC	LDL Class	Free Cholesterol	mg/dL
HDFC	HDL Class	Free Cholesterol	mg/dL
VLPL	VLDL Class	Phospholipids	mg/dL
IDPL	IDL Class	Phospholipids	mg/dL
LDPL	LDL Class	Phospholipids	mg/dL
HDPL	HDL Class	Phospholipids	mg/dL
HDA1	HDL Class	Apolipoprotein-A1	mg/dL
HDA2	HDL Class	Apolipoprotein-A2	mg/dL
VLAB	VLDL Class	Apolipoprotein-B100	mg/dL
IDAB	IDL Class	Apolipoprotein-B100	mg/dL

LDAB	LDL Class	Apolipoprotein-B100	mg/dL
V1TG	VLDL-1 Subclass	Triglycerides	mg/dL
V2TG	VLDL-2 Subclass	Triglycerides	mg/dL
V3TG	VLDL-3 Subclass	Triglycerides	mg/dL
V4TG	VLDL-4 Subclass	Triglycerides	mg/dL
V5TG	VLDL-5 Subclass	Triglycerides	mg/dL
V1CH	VLDL-1 Subclass	Cholesterol	mg/dL
V2CH	VLDL-2 Subclass	Cholesterol	mg/dL
V3CH	VLDL-3 Subclass	Cholesterol	mg/dL
V4CH	VLDL-4 Subclass	Cholesterol	mg/dL
V5CH	VLDL-5 Subclass	Cholesterol	mg/dL
V1FC	VLDL-1 Subclass	Free Cholesterol	mg/dL
V2FC	VLDL-2 Subclass	Free Cholesterol	mg/dL
V3FC	VLDL-3 Subclass	Free Cholesterol	mg/dL
V4FC	VLDL-4 Subclass	Free Cholesterol	mg/dL
V5FC	VLDL-5 Subclass	Free Cholesterol	mg/dL
V1PL	VLDL-1 Subclass	Phospholipids	mg/dL
V2PL	VLDL-2 Subclass	Phospholipids	mg/dL
V3PL	VLDL-3 Subclass	Phospholipids	mg/dL
V4PL	VLDL-4 Subclass	Phospholipids	mg/dL
V5PL	VLDL-5 Subclass	Phospholipids	mg/dL
L1TG	LDL-1 Subclass	Triglycerides	mg/dL
L2TG	LDL-2 Subclass	Triglycerides	mg/dL
L3TG	LDL-3 Subclass	Triglycerides	mg/dL
L4TG	LDL-4 Subclass	Triglycerides	mg/dL
L5TG	LDL-5 Subclass	Triglycerides	mg/dL
L6TG	LDL-6 Subclass	Triglycerides	mg/dL
L1CH	LDL-1 Subclass	Cholesterol	mg/dL
L2CH	LDL-2 Subclass	Cholesterol	mg/dL
L3CH	LDL-3 Subclass	Cholesterol	mg/dL
L4CH	LDL-4 Subclass	Cholesterol	mg/dL
L5CH	LDL-5 Subclass	Cholesterol	mg/dL
L6CH	LDL-6 Subclass	Cholesterol	mg/dL
L1FC	LDL-1 Subclass	Free Cholesterol	mg/dL
L2FC	LDL-2 Subclass	Free Cholesterol	mg/dL
L3FC	LDL-3 Subclass	Free Cholesterol	mg/dL
L4FC	LDL-4 Subclass	Free Cholesterol	mg/dL
L5FC	LDL-5 Subclass	Free Cholesterol	mg/dL
L6FC	LDL-6 Subclass	Free Cholesterol	mg/dL
L1PL	LDL-1 Subclass	Phospholipids	mg/dL
L2PL	LDL-2 Subclass	Phospholipids	mg/dL
L3PL	LDL-3 Subclass	Phospholipids	mg/dL
L4PL	LDL-4 Subclass	Phospholipids	mg/dL
L5PL	LDL-5 Subclass	Phospholipids	mg/dL
L6PL	LDL-6 Subclass	Phospholipids	mg/dL
L1AB	LDL-1 Subclass	Apolipoprotein-B100	mg/dL

L2AB	LDL-2 Subclass	Apolipoprotein-B100	mg/dL
L3AB	LDL-3 Subclass	Apolipoprotein-B100	mg/dL
L4AB	LDL-4 Subclass	Apolipoprotein-B100	mg/dL
L5AB	LDL-5 Subclass	Apolipoprotein-B100	mg/dL
L6AB	LDL-6 Subclass	Apolipoprotein-B100	mg/dL
H1TG	HDL-1 Subclass	Triglycerides	mg/dL
H2TG	HDL-2 Subclass	Triglycerides	mg/dL
H3TG	HDL-3 Subclass	Triglycerides	mg/dL
H4TG	HDL-4 Subclass	Triglycerides	mg/dL
H1CH	HDL-1 Subclass	Cholesterol	mg/dL
H2CH	HDL-2 Subclass	Cholesterol	mg/dL
H3CH	HDL-3 Subclass	Cholesterol	mg/dL
H4CH	HDL-4 Subclass	Cholesterol	mg/dL
H1FC	HDL-1 Subclass	Free Cholesterol	mg/dL
H2FC	HDL-2 Subclass	Free Cholesterol	mg/dL
H3FC	HDL-3 Subclass	Free Cholesterol	mg/dL
H4FC	HDL-4 Subclass	Free Cholesterol	mg/dL
H1PL	HDL-1 Subclass	Phospholipids	mg/dL
H2PL	HDL-2 Subclass	Phospholipids	mg/dL
H3PL	HDL-3 Subclass	Phospholipids	mg/dL
H4PL	HDL-4 Subclass	Phospholipids	mg/dL
H1A1	HDL-1 Subclass	Apolipoprotein-A1	mg/dL
H2A1	HDL-2 Subclass	Apolipoprotein-A1	mg/dL
H3A1	HDL-3 Subclass	Apolipoprotein-A1	mg/dL
H4A1	HDL-4 Subclass	Apolipoprotein-A1	mg/dL
H1A2	HDL-1 Subclass	Apolipoprotein-A2	mg/dL
H2A2	HDL-2 Subclass	Apolipoprotein-A2	mg/dL
H3A2	HDL-3 Subclass	Apolipoprotein-A2	mg/dL
H4A2	HDL-4 Subclass	Apolipoprotein-A2	mg/dL

Table S4. Demographic data for the training set

	SARS-Cov2 training set (n=7)	**Healthy Controls training set (n=8)**
Male Sex	4 (57.1%)	3 [37.5%]
Age, yrs [SD]	68.5 [±14.5]	60.3 [± 13.6]
BMI, kg/m^2 [SD]	31.9 [±6.4]	26.6 [± 3.9]
Diabetes (T1/T2)	0	0
Hypertension	3 [42.8%]	2 (20%)

Asthma	1 [14.2%]	0
COPD*	0	0
Arthritis	1 [14.2%]	0
Glaucoma	1 [14.2%]	0
Dyslipidaemia	0	0
Chronic Renal Disease	1 [14.2%]	0
Chronic Heart Disease	0	0

* Chronic obstructive pulmonary disease

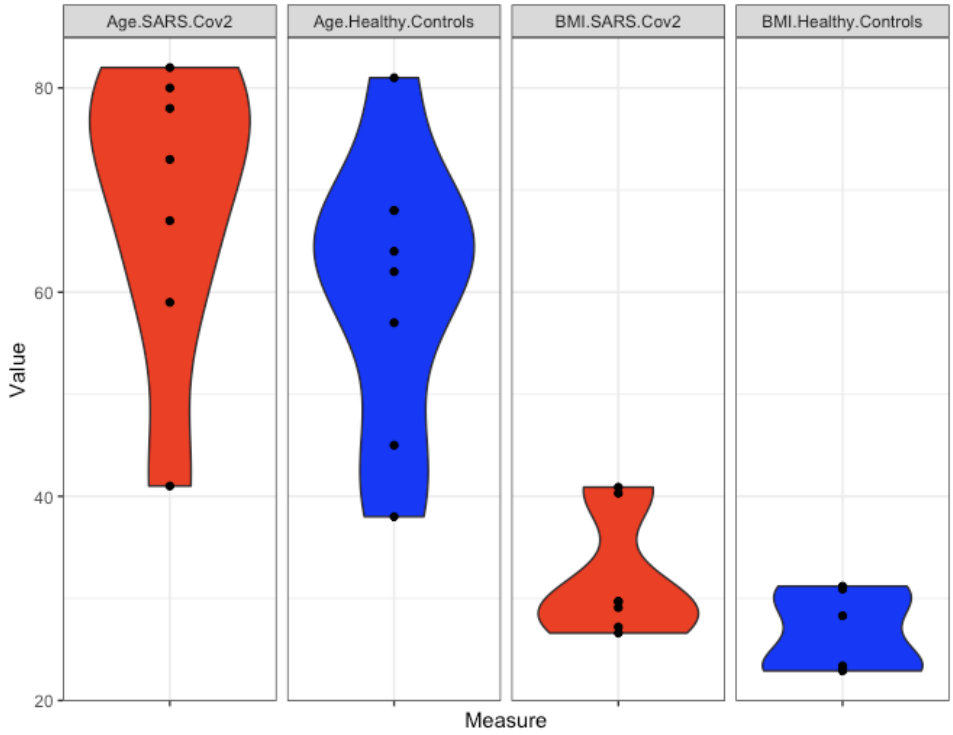

Figure S1. Violin plots presenting age and BMI of the patients in the matched training set

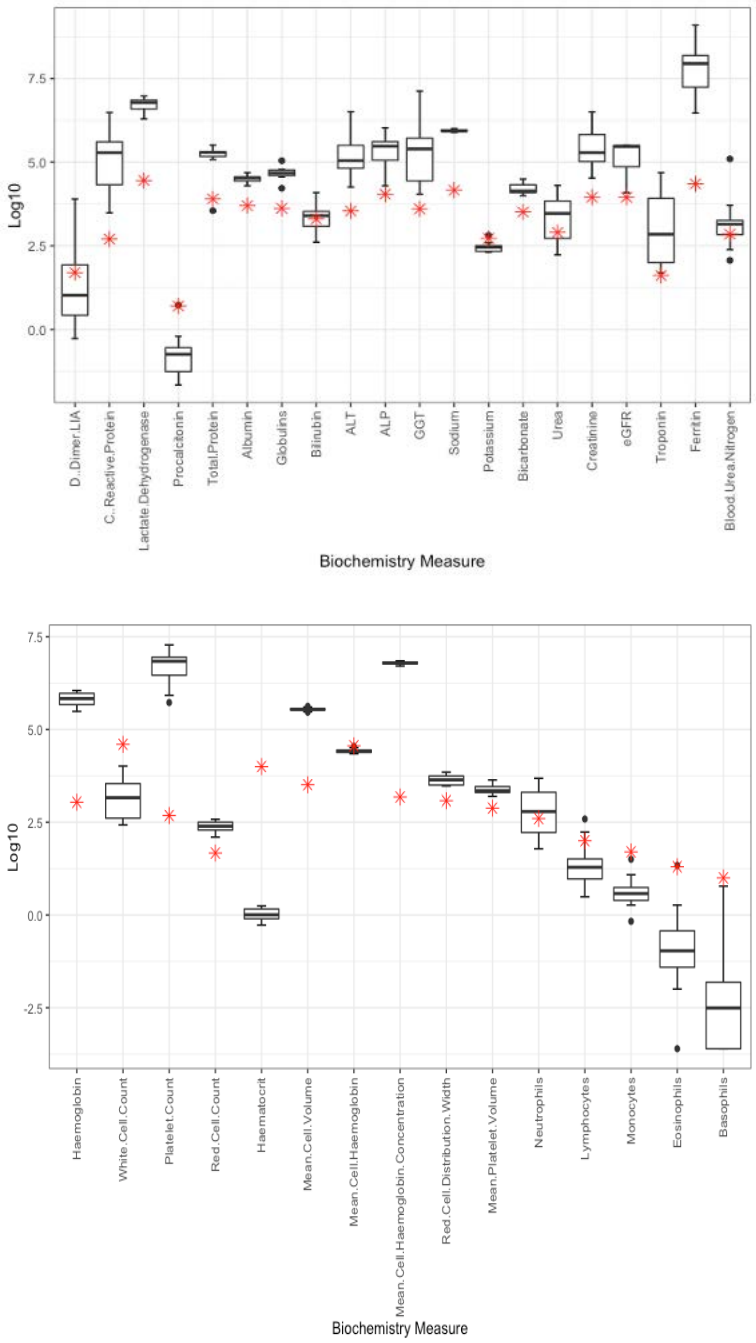

Figure S2. Blood biochemistry results from SARS-CoV-2 patients corresponding to timepoint 1. Asterix indicates upper limit of reference of pathology laboratory
Key: ALT – Alanine transaminase, ALP -Alkaline phosphatase, GGT – Gamma-glutamyl transferase, eGFR – estimated glomerular filtration rate.