# An Integrated Analysis and Comparison of Serum, Saliva and Sebum for COVID-19 Metabolomics

**Matt Spick**
University of Surrey
**Holly M Lewis**
University of Surrey
**Cecile F Frampas**
University of Surrey
**Katie Longman**
University of Surrey
**Catia Costa**
University of Surrey
**Alexander Stewart**
University of Surrey
**Deborah Dunn-Walters**
University of Surrey
**Debra J Skene**
University of Surrey
**Perdita Barran**
University of Manchester
**Michael J Wilde**
University of Leicester
**George Evetts**
Frimley Park Hospital NHS Foundation Trust
**Danni Greener**
Frimley Park Hospital NHS Foundation Trust
**Eleanor Sinclair**
University of Manchester
**Melanie J Bailey** (✉ m.bailey@surrey.ac.uk)
University of Surrey

**Research Article**

**Keywords:** Sampling Matrices, COVID-19, Metabolomics, Lipidomics,Mass Spectrometry

**Posted Date:** March 1st, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1337471/v1



**Version of Record:** A version of this preprint was published at Scientific Reports on July 13th, 2022. See the published version at https://doi.org/10.1038/s41598-022-16123-4.

# Abstract

Non-invasive sampling approaches are desirable in healthcare settings due to the lower burden on patients and clinical staff. In the field of metabolomics, however, most studies have utilised blood serum or plasma. In this work 83 COVID-19 positive and negative hospitalised participants provided serum, saliva and sebum samples for analysis by mass spectrometry. Here we present the first comprehensive analysis of correlations between serum metabolites, salivary metabolites and sebum lipids, and consider their relative accuracy in differentiating COVID-19 positive participants from controls. Sebum lipids showed clear correlations to serum metabolites, with widespread changes to the serum-sebum lipid axis in COVID-19 positive participants, evidence of multi-organ dyslipidemia. In the COVID-19 positive cohort, correlations were notably marked between sebum lipids and the immunostimulatory hormone dehydroepiandrosterone sulphate. In terms of diagnostic accuracy, serum performed best by multivariate analysis (sensitivity and specificity of 0.97), with the dominant changes in triglyceride and bile acid levels, concordant with other studies identifying dyslipidemia as a hallmark of COVID-19 infection. Sebum diagnostic accuracy performed well (sensitivity 0.92; specificity 0.84), with saliva offering weaker diagnostic accuracy (sensitivity 0.78; specificity 0.83). These findings highlight the potential for integrated biofluid analyses to provide insight into the whole-body atlas of pathophysiological conditions.


# 1. Introduction

Since 2020, the COVID-19 pandemic has been at the forefront of global health. Whilst great strides have been made in fields as diverse as diagnosis, drug therapy and vaccination[1,2], SARS-CoV-2 still poses a major health risk to the world. This reflects both the difficulty of eradicating all disease reservoirs, and also the potential for the virus to evolve new variants of concern[3], potentially leading to endemic waves of infection or vaccine escape[4]. Whilst the symptoms of COVID-19 are now well-described, the metabolic processes underlying these changes are incompletely understood. Additionally, as supply chains for testing and related consumables have collapsed at the height of recent waves, alternative approaches to diagnostics, and indeed prognostics, will be critical for future epidemics. Metabolomics offers one such alternative, but reflecting the urgency of conducting research in a pandemic, most metabolomics research to date has focused on biomarker discovery rather than on validation[5]. The vast majority of biomarker discovery research for COVID-19 and other diseases has been carried out using blood as a sampling matrix. As a biofluid rich in metabolites not prone to external contamination, blood sampling offers considerable advantages in terms of biological interpretation, offering diagnostic power via clinical indicators such as lymphocytes as well as metabolites[6,7]. However, collection of blood is onerous and invasive, both for clinics and patients. In contrast, saliva and sebum (a swab from the back of the neck) can be collected quickly and painlessly. This puts a lower burden on both clinicians and patients, especially for young children and older people.

Metabolomic biomarkers identified in different COVID-19 studies have naturally varied according to extraction and analytical techniques. [8] Several blood metabolomics studies have identified dysregulation

of amino acid levels or ratios thereof, for example the kynurenine / arginine ratio has been suggested as a biomarker of COVID-19[9,10], as has the kynurenine / tryptophan ratio[11 12]. Glutamine has also been proposed as a biomarker, both in its own right[13,14], and as a ratio with glutamate[11]. Dyslipidemia has been identified as a hallmark of COVID-19 infection, with increased triglyceride levels a common observation[15,16,11]; phosphatidylcholines have also been identified as potential biomarkers[12,17]. In contrast to analyses of blood, however, investigation of COVID-19 induced changes in the skin lipidome and the salivary metabolome have to date been few in number and restricted to untargeted mass spectrometry methods or infra-red spectroscopy studies that do not provide full identification of biomarkers[18,19,20].

In this study, correlation analysis was performed between serum, sebum and saliva features to provide a more holistic view of COVID-19 related dysregulation. Integrated analyses incorporating different ‘omics workflows have greater ability to capture biological complexity and have helped the understanding of COVID and other illnesses[21,22]. In addition, dermatological symptoms of COVID-19 have been reported[23], including vesicular, urticarial and vascular rashes[24]. To our knowledge, no integrated analysis of sebum, saliva and serum for COVID-19 or other illnesses has yet been performed, and this study aims to fill this gap.

For metabolomic diagnosis of COVID-19, blood-based analyses have outperformed less-invasive analysis such as breath, sebum or saliva, with one meta-analysis finding that blood-based analyses delivered average sensitivity of 0.89 and specificity of 0.96, compared with 0.76 and 0.81 across the less invasive biofluids[5]. Meta-analysis of independent studies does not, however, take account of between-study confounders. In addition to analyses of biofluid feature correlations, therefore, diagnostic accuracy was measured for serum, sebum and saliva, allowing for direct comparisons of diagnostic power.

# 2. Materials And Methods

# 2.1 Participant recruitment and ethics

### Ethical approval

for this project (IRAS project ID 155921) was obtained via the NHS Health Research Authority (REC reference: 14/LO/1221). The participants included in this study were recruited consecutively (i.e. without selective sampling criteria other than suspicion of COVID-19 infection) at Frimley Park NHS Trust, totalling 83 participants, between May 2020 and July 2020. Participants were identified by clinical staff to ensure that they had the capacity to consent to the study and were asked to sign an Informed Consent Form; those that did not have this capacity were not sampled. At the time of recruitment, participants were categorised by the hospital as either “query COVID” (meaning there was clinical suspicion of COVID-19 infection, but a negative positive RT-PCR SARS-CoV-2 test result had been recorded during their admission) or “COVID positive” (meaning that a positive test result had been recorded). All participants

were provided with a Patient Information Sheet explaining the goals of the study. All methods performed as part of this study were performed in accordance with the relevant guidelines and regulations.

## 2.2 Sample collection, extraction and processing

Collection of the samples was performed by researchers from the University of Surrey at Frimley Park NHS Foundation Trust hospitals; participants were requested to provide all three biofluids, but due to declined consent for blood sampling, or inability to express saliva or easily provide blood, not all participants provided all three samples (Table S1). Serum collection and extraction followed the protocols set out by the COVID-19 Coalition. [36] In brief, venous blood was collected in 3 mL serum tubes, transported to University of Surrey by courier whilst stored on ice, and centrifuged on arrival at 1,600 g for 10 mins at 4°C. All samples with a sampling time interval greater than four hours were rejected. Serum was then decanted into 100 µl aliquots and stored at -80°C until processing. Prior to analysis, the serum was sterilised using 200 ul of ethanol into 100 µl of serum (2:1 v/v solvent/sample ratio). Saliva extraction was performed as described in Frampas *et al*, [18] again following the protocols established by the COVID-19 MS Coalition. [36] All samples (sebum, saliva and serum) were taken from each patient within 20 minutes of one another.

Sebum collection, extraction and processing was performed as described in Spick *et al*, [19] using an untargeted liquid-chromatography mass spectrometry (LC-MS) methodology. Features with a coefficient of variation (CV%) across all pooled QCs above 20% were removed, as were those that were not present in at least 90% of pooled QC injections. These features were then field blank adjusted: all features with a signal to noise ratio below 3x were also rejected. The remaining set of 998 features were deemed to be robust and reproducible.

Serum samples were analysed using the Biocrates MxP Quant 500 system using a Xevo TQ-S Triple Quadrupole Mass Spectrometer coupled to an Acquity UPLC system (Waters Corporation, Milford, MA, USA). The MxP Quant 500 system provides targeted quantification of metabolites including amino acids and derivatives, bile acids, biogenic amines, acylcarnitines, carbohydrates and other small molecule metabolites, plus a wide array of lipids. Analysis takes place via a single assay, and two analytical procedures. The first of these procedures operates by liquid chromatography (operated in both positive and negative ion mode) and the second by flow injection analysis (positive ion mode), both coupled to tandem mass spectrometry with isotopically labelled internal standards for quantification. Sample order was randomised across 96-well plates, and 3 levels of quality controls (QC) were run on each plate. Blank PBS (phosphate-buffered saline) samples (three technical replicates) were used for the calculation of the limits of detection (LOD). Biogenic amines and amino acids were quantified for each plate using a seven-point calibration curve, with other analytes semi-quantitated with a single point standard (i.e. assuming concentration linearity in the range measured). The levels of metabolites present in each QC were compared to the expected values and the CV% calculated. Data were normalised between the three batches using the results of quality control level 2 (QC2) repeats across the plate (n = 5) and between plates (n = 3) using Biocrates METIDQ software (QC2 correction). Metabolites where > 25%

concentrations were at or below the limit of detection (≪LOD), above the limit of quantification (> LOQ), or where the blank was out of range were excluded (total n excluded in serum = 150). The remaining 474 quantified metabolites comprised of 8 acylcarnitines, 20 amino acids, 26 biogenic amines, 11 bile acids, 53 ceramides, 15 cholesteryl esters, 1 cresol, 9 diglycerides, 4 carboxylic and fatty acids, 85 phosphatidyl cholines, 14 sphingolipids, 222 triglycerides, 2 hormones, 2 indoles, 1 nucleobase and 1 vitamin.

Saliva samples were quantified in a similar way to the serum samples but employing the Biocrates AbsoluteIDQ p180 system, which can detect up to 188 metabolites, fewer than the MxP Quant 500. The method used was as described for the MxP Quant 500 system, i.e. the sample order was randomised, the same 3 levels of QCs were run on each 96-well plate, and the same protocols were followed to establish CVs, LODs and LOQs. As for serum, metabolites where > 25% concentrations were at or below LLOQ, where measurements were above LOQ, or where the blank was out of range were excluded (total n excluded in saliva = 103). The remaining 83 quantified metabolites comprised 7 acylcarnitines, 18 amino acids, 5 biogenic amines, 42 glycerophospholipids and 11 sphingolipids.

In this study identifications were made in accordance with the Metabolomics Standards Initiative for metabolite identification[37]. Serum and saliva metabolites were identified and quantified using isotopically labelled internal standards, retention times and multiple reaction monitoring using optimised mass spectrometry conditions as provided by Biocrates. Sebum features in this work were either identified as putatively annotated compounds based on accurate *m/z* matching, or were unknown compounds with differentiable and quantifiable MS signals but no library identification.

For all three biofluids, the output was a data block in the form of a peak:area matrix of *n* participants by *p* features.

# 2.3 Statistical analysis

Initial pre-processing of the three data blocks was conducted in Progenesis QI (for sebum samples) and using manufacturer software (for the Biocrates kits) as described in the preceding section. Following this, missing value replacement was carried out using K-nearest neighbours; all data were then log transformed, mean centred and pareto scaled using the web-based platform MetaboAnalyst[38]. Analysis of the processed peak:area matrices was then conducted in MetaboAnalyst together with user-written scripts in the programming language Python and the package scikit-learn[39,40]. Participant characteristics (positive and negative cohorts) were assessed by two-tailed t-tests, or by two-tailed Mann Whitney U tests where parameters were not normally distributed.

To explore relationships between the different sampling matrices, correlation analysis by Pearson correlation coefficient was conducted across the data blocks, for matched metabolites between saliva and serum, and for the overall matrices for sebum and serum / sebum and saliva (as matched analysis was not possible due to no overlap between these biofluids).

To compare the ability of the different sampling matrices to differentiate between positive and negative COVID-19 participants, logistic regression using recursive feature elimination with cross validation was performed in Python using the scikit-learn package for each block, inclusive of metabolic ratios previously identified as diagnostic (serum and saliva ratios of kynurenine / arginine, kynurenine / tryptophan and glutamine / glutamate). Recursive feature elimination was performed to limit the feature set employed and reduce overfitting[41], leading to a set of smaller data blocks to be analysed. Partial least squares-discriminant analysis (PLS-DA) was then conducted for each of the three data blocks (serum, sebum and saliva) using the reduced feature set.

Leave-one-out cross-validation (LOOCV) was used for model validation to test accuracy, sensitivity and specificity using COVID-19 RT-PCR results as the ground truth. Sensitivity was defined as the true positive rate, i.e. the probability that a positive test result will be obtained when the disease is present, and calculated from a confusion matrix as true positives / (true positives plus false negatives). Specificity was defined as the true negative rate, i.e. the probability that a negative test result will be obtained when the disease is not present, and calculated as true negatives / (true negatives plus false positives). Variable importance in projection (VIP) scores were used to assess feature / metabolite significance.

# 3. Results

## 3.1 Population metadata overview

The study population analysed in this work included 83 participants, comprising 40 participants presenting with a positive COVID-19 RT-PCR test and 43 participants with a negative RT-PCR test but presenting with similar clinical symptoms to COVID-19. A summary of the metadata is shown in Table 1. The populations providing sebum, saliva and serum samples each represented a subset of this group, as not all participants consented to provide all biofluids. Complete data by subgroup is shown in Supplementary Material, Table S1. The variation in biofluids sampled was mainly due to either participants not consenting to give blood on the grounds of invasiveness of the sampling, or participants being unable to generate sufficient saliva or blood.

Table 1
Characteristics of study population

| Parameters | Negative for COVID-19 | Positive for COVID-19 | p-value |
|---|---|---|---|
| N | 43 | 40 | |
| Age (mean, standard deviation; years) | 62.9 ± 19.2 | 61.4 ± 19.8 | 0.74 |
| Male / Female (n) | 20 / 23 | 20 / 20 | 0.54 |
| Treated for Hypertension (n) | 18 | 13 | 0.39 |
| Treated for High Cholesterol (n) | 10 | 6 | 0.36 |
| Treated for Type 2 Diabetes Mellitus (n) | 12 | 11 | 0.97 |
| Treated for Ischaemic Heart Disease (n) | 7 | 5 | 0.64 |
| Current Smoker (n) | 2 | 2 | 0.94 |
| Ex-Smoker (n) | 13 | 6 | 0.11 |
| Medical Acute Dependency admission (n) | 7 | 13 | 0.09 |
| Intensive Care Unit admission (n) | 1 | 5 | 0.10 |
| Survived Admission (n) | 41 | 37 | 0.62 |
| Duration of pre-admission symptoms (mean, standard deviation; days) | 11.9 ± 20.2 | 6.7 ± 6.8 | 0.12 |
| Time between positive RT-PCR test and sampling (mean, standard deviation; days) | NA | 5 ± 7 | |
| Lymphocytes (mean, standard deviation; cells / μL) | 1.0 ± 0.5 | 0.7 ± 0.3 | 0.002 |
| C-Reactive Protein (mean, standard deviation; mg / L) | 138.4 ± 96.4 | 170.8 ± 121.2 | 0.20 |
| Eosinophils (mean, standard deviation; 100 / μL) | 0.3 ± 0.4 | 0.2 ± 0.3 | 0.007 |
| Bilateral Chest X-Ray changes (n) | 6 | 22 | 0.001 |
| Continuous Positive Airway Pressure (n) | 4 | 10 | 0.07 |
| O2 required (n) | 14 | 21 | 0.07 |

Age and sex distributions for COVID-19 positive and negative cohorts were similar (mean age of 61.4 years and 62.9 years, and M:F ratios of 1.00x and 0.87x respectively). On average, participants had seen 8 days of pre-admission symptoms. Comorbidities are associated with both hospitalisation and more severe outcomes for COVID-19 infection, but will also alter the metabolome of participants, representing both a causative and confounding factor. Due to hospital recruitment, however, comorbidities including T2DM, hypertension, high cholesterol and ischaemic heart disease were represented in both the positive and negative groups. Ex-smokers and current smokers were more highly represented in the COVID-19

negative group (35% of the negative participants, versus 20% of the positive participants). All participants had at least a clinical suspicion of COVID-19 infection, thus respiratory distress due to past or present smoking may have caused ‘over-recruitment’ of smokers in the cohort that subsequently tested negative.

Levels of C-Reactive Protein (CRP) were higher for COVID-19 positive participants, whilst lymphocyte and eosinophils levels were lower. A two-tailed Mann Whitney U test on CRP levels, lymphocytes and eosinophils provided p-values of 0.20, 0.002 and 0.007, respectively. Effect sizes (calculated by Cohen’s D) were 0.28, 0.39 and 0.34, respectively. COVID19 positive participants were also more likely to present with bilateral chest X-ray changes, more likely to require oxygen / CPAP, and were also escalated to ICU and MADU more frequently. These observations were in agreement with literature descriptions of COVID-19 symptoms, clinical indicators and progression[25].

# 3.2 Feature identification

For serum and saliva, metabolite identification was performed under manufacturer protocol using internal standards together with accurate mass matching, also allowing for quantification of concentrations. For saliva the AbsoluteIDQ p180 system generated 83 identified metabolites that were reliably quantified (out of a theoretical maximum of 188). For serum the MxP Quant 500 system generated 472 identified metabolites that were reliably quantified in samples (out of a theoretical maximum of 630). For sebum, an untargeted lipidomics approach was used, with accurate mass matching using the Progenesis QI software, with tandem MS where possible. A total of 998 sebum features (both putatively identified lipids and unidentified compounds) were considered to be reliable and robust, as set out in Spick *et al*. [19] A limited metabolite set was also investigated for serum to provide an equivalent comparison of diagnostic accuracy for serum versus saliva (i.e. to remove the advantage of the serum peak:area matrix being more feature-rich than saliva).

# 3.3 Correlation analysis: serum and sebum lipids

Whilst sebum was analysed via an untargeted LC/MS method, and so the putative features identified cannot be matched against direct serum metabolites, correlations between the sebum and serum peak:area matrices were investigated. Figure 1 shows a heatmap of Pearson correlation coefficients for a subset of metabolites and lipids showing the strongest relationships between the sebum and serum data blocks, separated between COVID-19 negative (Fig. 1A) and positive (Fig. 1B) participants.

Correlation coefficients in Fig. 1 show clustering and some of the observed correlations are relatively high, reaching 0.6. As can be seen in the lower part of Fig. 1A for COVID-19 negative participants, several serum carnitines (especially hexadecenoylcarnitine) and L-Proline Betaine correlate negatively with sebum lipids. There is also a set of general and relatively strong positive correlation coefficients (0.5 to 0.6) between serum triglycerides and several sebum glycerides, visible on the left side of Fig. 1A. In the cohort of COVID-19 positive participants, however, this general pattern breaks down (Fig. 1B), and different serum metabolites show increased / decreased correlation with sebum glycerides. Figure 2 illustrates the positive correlation between serum DHEAS (dehydroepiandrosterone sulphate) and three

specific diglycerides in sebum for COVID-19 positive participants; these correlations were markedly weaker in COVID-19 negative participants.

## 3.4 Correlation analysis: serum and saliva

Both serum and saliva samples were processed using a standardised Biocrates platform employing internal standards for accurate quantification of concentrations. This allowed investigation of the extent of correlation of salivary and serum concentrations for 79 identified metabolites for the same patients, both directly (correlation of specific metabolites in serum versus saliva) and generally (overall correlation between the two data blocks). Figure 3 illustrates the correlations across all 79 metabolites. The average Pearson correlation coefficient for paired metabolites (e.g. taurine in serum versus taurine in saliva) was negligible for both COVID-19 positive and negative participants; only one paired correlation coefficient was outside the range of 0.2 to -0.2, this exception being trans-4-hydroxyproline (t4-OH-Pro). Some serum metabolites showed a general pattern of positive correlation in the COVID-19 controls such as serum leucine and isoleucine being generally associated with higher concentrations of salivary metabolites (Fig. 3A). These positive correlations were reduced in COVID-19 positive participants (Fig. 3B).

## 3.5 Correlation analysis: serum and sebum

Finally, for completeness, correlations between salivary metabolites and sebum lipids were also investigated. Similar results are observed for correlations between sebum and saliva, as was the case for serum and saliva. In the controls (COVID-19 negative participants) correlations are relatively low between salivary metabolites and serum lipids, with some positive relationships between salivary amino acids and some glycerides (Fig. 4A). In COVID-19 positive participants, a series of negative correlations become evident between salivary phosphatidylcholines and a series of sebum lipids (Fig. 4B).

## 3.6 Diagnostic performance

To further investigate dysregulation due to COVID-19 between biofluids, the ability of each of the biofluids to differentiate between positive and negative cases was analysed by PLS-DA analysis of their respective reduced feature data blocks, and sensitivity / specificity / accuracy for each data set was calculated together with confidence intervals. Two models were constructed for serum – one with the full Biocrates MxP Quant 500 metabolite set, and one using a metabolite set equivalent to the p180 Biocrates system to allow for a more comparable assessment versus the saliva PLS-DA model.

Table 2
Comparison of PLS-DA model performance using LOOCV

| Biofluid | *n* patients (pos / neg) | *p* features / after RFECV[b] | Sensitivity | Specificity | Accuracy [a] | R2Y | Q2Y |
|---|---|---|---|---|---|---|---|
| Saliva | 47 (23 / 24) | 83 / 5 | 0.78 | 0.83 | 0.80 | 0.42 | 0.26 |
| Sebum | 80 (37 / 43) | 998 / 26 | 0.92 | 0.84 | 0.88 | 0.63 | 0.51 |
| Serum | 63 (30 / 33) | 472 / 41 | 0.97 | 0.97 | 0.97 | 0.90 | 0.72 |
| Serum (p180 only) | 63 (30 / 33) | 86 / 23 | 0.83 | 0.94 | 0.89 | 0.78 | 0.47 |

[a] All PLS-DA models used 5 components except saliva which found maximum accuracy with 3 components

[b] Number of features in model is the smaller number after Recursive Feature Elimination with Cross-Validation

As can be seen from Table 2, the best diagnostic performance was delivered by serum using a reduced panel of 41 metabolites, with sensitivity of 0.97 (95% confidence interval of 0.83–1.00) and specificity of 0.97 (0.84–1.00). Restricting the serum feature set to metabolites detected by the p180 Biocrates kit lowered sensitivity to 0.83 (0.65–0.94) and specificity to 0.94 (0.80–0.99). Sebum saw a major reduction in features using RFECV, from 998 to 26, and the resulting PLS-DA model achieved sensitivity of 0.92 (95% confidence interval of 0.78–0.98) and specificity of 0.84 (0.69–0.93). Saliva delivered a PLS-DA model with sensitivity of 0.78 (95% confidence interval of 0.56–0.93) and specificity of 0.83 (0.63–0.95), performing less well than the other biofluids. The reduced data sets were also investigated for confounders by principal component analyses; neither age nor sex showed clustering for serum, saliva or sebum (Figures S2 to S4, Supplementary Material).

The datasets were additionally processed by support vector machine (SVM), logistic regression and random forest models, and in all cases, comparable sensitivity and specificity was achieved (Table S2, Supplementary Material), confirming that separation is not sensitive to the model used. Figure 5 shows separation of COVID-19 positive and negatives for serum by PLS-DA as well as the highest VIP score metabolites.

Separation for serum can be seen visually in Fig. 5A with two components, rising to 0.97 accuracy with 5 components. The high VIP score metabolites are dominated by lipids, especially triglycerides, but the highest VIP metabolite was glycolithocholic acid 3-sulfate (GLCAS), a conjugated bile acid. Of the amino acids, L-proline betaine had the highest VIP score. Separation by PLS-DA is also shown below for both sebum (Fig. 6) and saliva (Fig. 7). Separation becomes visually worse in the order serum / sebum /

saliva, with saliva also featuring a notably small number of high VIP metabolites, reflecting the difficulty of RFECV finding a local maximum for accuracy with a noisy dataset.

Finally, whilst for correlation analysis and for assessing important metabolites / classification, the largest possible populations have been used, it is also important to compare diagnostic accuracy for the exactly matched 37 participants for which all three samples were taken (Figure S1, Supplementary Material). For this subset of 37 participants, diagnostic accuracy for serum, sebum and saliva (measured by LOOCV) was 0.95, 0.87 and 0.70, the same ranking as shown in Table 2 and almost identical classification accuracy.

# 4. Discussion

In this study sebum showed strong correlation with serum metabolites, revealing dysregulation across biofluids. Notably, in the control group of COVID-19 negative participants, a set of positive correlations between serum triglycerides and ceramides and skin lipids was visible. This positive correlation in the controls was dysregulated in the cohort of COVID-19 positive participants, evidence of multi-organ dyslipidemia due to COVID-19. The integrated analysis presented here also showed correlation between sebum lipids and DHEAS in the cohort of COVID-19 positive participants. DHEAS is an immune-system positive adrenal hormone and also an antiglucocorticoid, and so alterations in any DHEAS / sebum axis could be indicative of immune response, and may underpin diagnostic differences seen in the sebum of COVID-19 positive and negative participants[26,27]. Sebum lipids have been identified as biomarkers in other pathologies, such as Parkinson's Disease where sebum dysregulation has been linked to carnitines, a relationship that is also visible in this pathology[28,29]. Although as yet not as well investigated as blood-based metabolomics, we believe that sebomics holds promise for investigating other conditions.

Saliva showed weaker correlation to serum, especially in the case of directly matched metabolites. COVID-19 positivity did change the correlation maps between the two biofluids, but from a less correlated starting point than sebum / serum, and resulting in weaker diagnostic power overall. As a filtrate, saliva should be influenced by serum levels, but concentrations are lower than in blood[30]. Furthermore, the salivary biome is independent and has its own discrete functions and is markedly more subject to direct contamination from food or medication. Indeed, the correlation of metabolites between saliva and blood has previously been found to be weak or in some cases non-existent[31,32], and the results of this work are concordant with previous studies that show that the most accurate metabolomic analyses are those that target blood metabolites[5]. This is unsurprising, given that blood is homeostatically controlled in a way that saliva and sebum are not. Saliva and sebum showed weak correlation in the control (COVID-19 negative) group, but perhaps surprisingly, in the COVID-19 positive group a series of negative correlations between phosphatidylcholines and sebum lipids were visible. A direct causal relationship is challenging to identify given that saliva is a filtrate, but an indirect relationship is possible. One study found a relationship between reduced levels of phosphatidylcholines and Alzheimer's disease[33], but in general research on the metabolomics of saliva in areas other than oral health has been limited.

The analysis of diagnostic capability of a reduced feature set for each biofluid showed declining accuracy for the biofluids in the order serum (diagnostic accuracy 0.97), sebum (0.88), and saliva (0.80). Serum therefore performed best in this comparison of matched biofluids, but sebum also performed relatively well, and in this reanalysis better than previously reported[19]. This was in large part due to the use of feature reduction: the original sebum dataset's 998 features are likely to have led to overfitting (exceeding the number of samples by a multiple of 15), worse generalisation and worse performance on cross-validation. It should be stressed, however, that n for all three biofluids in this work was small, and so these results are indicative of relative performance only. Without a validation cohort the accuracies presented here should not be taken as indicative of absolute performance.

Diagnostic accuracy was also investigated for serum using a more limited set of metabolites, equivalent to that provided by the p180 Biocrates kit, leading to reduced accuracy but still relatively better than the other biofluids. This finding illustrates the trade-off between narrowly targeted analyses and widely targeted (or untargeted) analyses – whilst it is easier to validate a more tightly controlled panel of metabolites, a wider range can reveal additional biomarkers, especially during the initial discovery phase of biomarker identification. The biomarkers responsible for separation between positive and negative measured by VIP score were: glycolithocholic acid 3-sulfate (GLCAS), a bile acid, two triglycerides (TG(22:4_32:2) and TG(18:3_33:2)), as well as the amino acid L-proline betaine. This is consistent with other studies finding evidence of dyslipidemia, particularly increased triglyceride levels[15,16,11]. The dysregulation of GLCAS is also concordant with liver damage caused by COVID-19[34], and dysregulation of bile acids (deoxycholic acid and ursodeoxycholic / hyodeoxycholic acid) has previously been reported as a key feature specific to COVID-19, differentiating between COVID-19 and other respiratory and inflammatory diseases in hospital-recruited patients[35].

Saliva performed less well in diagnostic terms, in spite of evidence of pathology-driven correlations with other biofluids, showing weaker ability to differentiate COVID-19 negative from positive in this analysis. It should be noted, however, that due to the inability for ethical reasons to require abstinence from food or drink in the hospital setting, saliva would have been the most subject to environmental confounders such as the recent oral intake of food or medication. Whilst a clear limitation of the study, it does also reflect the practical limitations of sampling during a pandemic or indeed in any busy clinic.

Overall, whilst the integrated analysis herein of serum, sebum and saliva shows challenges in identifying reproducible metabolic biomarkers of COVID-19, it also shows the potential for non-invasive sampling in revealing relationships across biofluids and pathways. For diagnostic purposes where sensitivity and specificity are paramount, however, we believe that blood-based metabolomics will remain the best-in-class approach.

# Declarations

### Data sharing statement

All data relating to this work will be made available on the Zenodo data repository following publication. The analytical protocols used as well as mass spectrometry .raw files, sample and participant data will be openly available for all researchers to access.

### Competing Interest

Professor Barran reports patent PCT/GB2019/052169. The other authors have no interests to declare.

### Author's contributions

MS was responsible for statistical analysis and authorship of the manuscript, which was reviewed and edited by MB and DJS. MS and MW extracted and processed the sebum samples analysed herein, KL and CF extracted saliva samples, and HL extracted, processed and validated serum samples and also processed and validated saliva samples. KL, CF and AS collected all patient samples used in this work. AS and DDW obtained ethical approval. GE and DG facilitated access to participants and collected participant metadata. PB, DJSand CC assisted with mass spectrometry method development and MW advised on machine learning techniques. MB, DDW, DJSand PB obtained funding for the study; MB was responsible for supervision of the research team.

### Funding

The authors would like to acknowledge funding from the EPSRC Impact Acceleration Account for sample collection, as well as EPSRC Fellowship Funding EP/R031118/1. Mass Spectrometry was funded under EP/P001440/1. Sample collection and processing was funded by the University of Surrey and the BBSRC BB/T002212/1.

### Acknowledgements

The authors acknowledge the support of the COVID-19 International Mass Spectrometry (MS) Coalition.[42]In addition, the authors are grateful to Samiksha Ghimire from Groningen Medical School for translation of participant information sheets and consent forms into Nepalese, and to Thanuja Weerasinge (Jay), Manjula Meda, Chris Orchard and Joanne Zamani of Frimley Park NHS Foundation Trust for their help with ethics approvals and access to hospital patients.Jeroen Pennings provided advice on machine learning techniques. Graphical abstract made using BioRender (Biorender.com).

# References

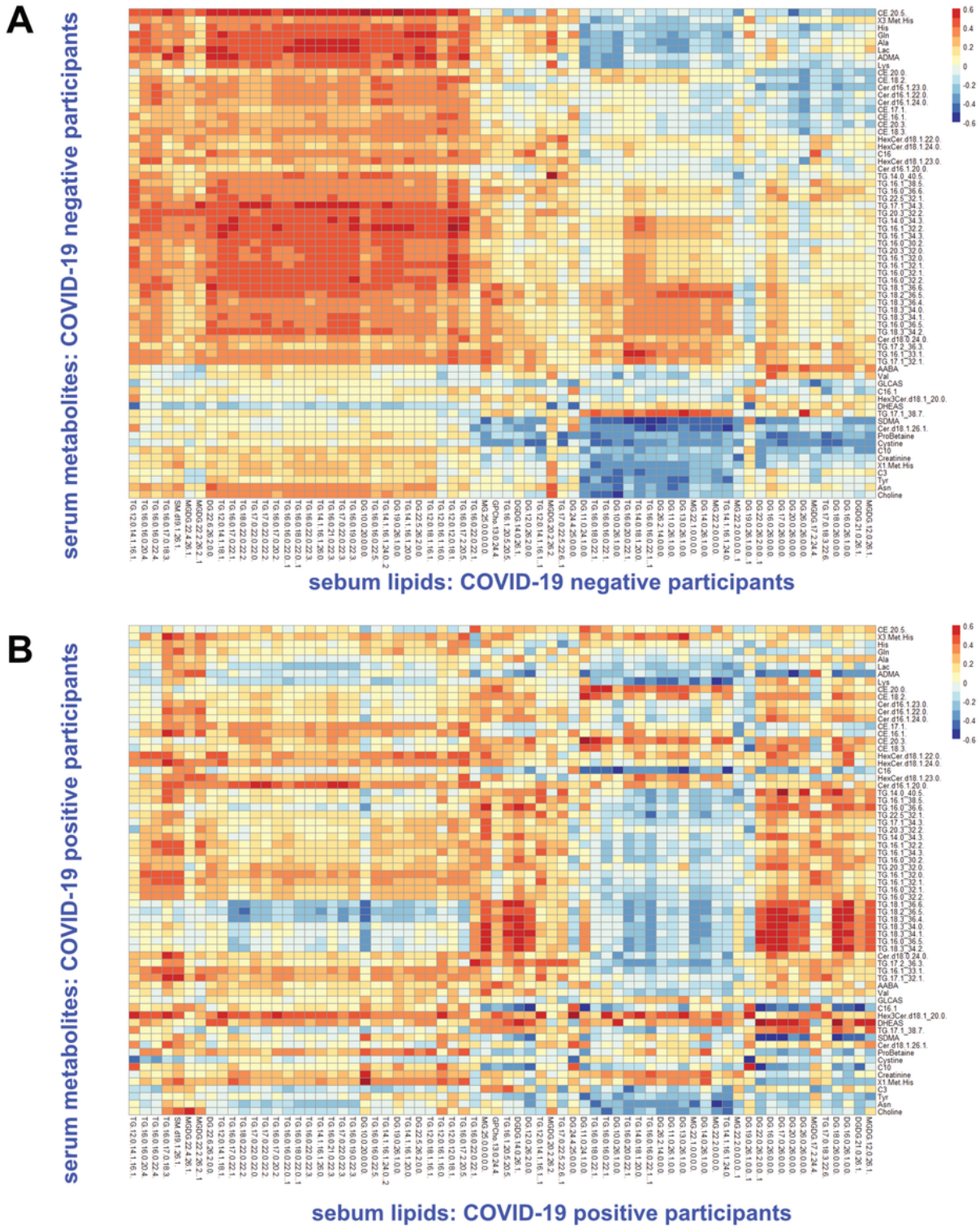
1. Knoll, M. D. & Wonodi, C. Oxford-AstraZeneca COVID-19 vaccine efficacy. *Lancet (London, England)***397**, 72–74 (2021).
2. The RECOVERY Collaborative Group. Dexamethasone in Hospitalized Patients with Covid-19 — Preliminary Report. *N. Engl. J. Med.* (2020). doi:10.1056/nejmoa2021436

3. WHO. Classification of omicron (b.1.1.529) sars-cov-2 variant-of-concern. (2021). Available at: https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern. (Accessed: 1st December 2021)
4. Rella, S. A., Kulikova, Y. A., Dermitzakis, E. T. & Kondrashov, F. A. Rates of SARS-CoV-2 transmission and vaccination impact the fate of vaccine-resistant strains. *Sci. Rep.***11**, 15729 (2021).
5. Spick, M. *et al.* Systematic review with meta-analysis of diagnostic test accuracy for COVID-19 by mass spectrometry. *Metabolism* 154922 (2021). doi:10.1016/J.METABOL.2021.154922
6. Kukar, M. *et al.* COVID-19 diagnosis by routine blood tests using machine learning. *Sci. Rep.***11**, 10738 (2021).
7. Barberis, E. *et al.* Understanding protection from SARS-CoV-2 using metabolomics. *Sci. Rep.***11**, 13796 (2021).
8. Pang, Z., Zhou, G., Chong, J. & Xia, J. Comprehensive Meta-Analysis of COVID-19 Global Metabolomics Datasets. *Metabolites***11**, (2021).
9. Thomas, T. *et al.* COVID-19 infection alters kynurenine and fatty acid metabolism, correlating with IL-6 levels and renal status. *JCI Insight***5**, (2020).
10. Fraser, D. D. *et al.* Metabolomics Profiling of Critically Ill Coronavirus Disease 2019 Patients: Identification of Diagnostic and Prognostic Biomarkers. *Crit. Care Explor.***2**, e0272 (2020).
11. Kimhofer, T. *et al.* Integrative Modeling of Quantitative Plasma Lipoprotein, Metabolic, and Amino Acid Data Reveals a Multiorgan Pathological Signature of SARS-CoV-2 Infection. *J. Proteome Res.***19**, 4442–4454 (2020).
12. López-Hernández, Y. *et al.* Targeted metabolomics identifies high performing diagnostic and prognostic biomarkers for COVID-19. *Sci. Rep.***11**, 14732 (2021).
13. Doğan, H. O. *et al.* Understanding the pathophysiological changes via untargeted metabolomics in COVID-19 patients. *J. Med. Virol.***93**, 2340–2349 (2021).
14. Shen, B. *et al.* Article Proteomic and Metabolomic Characterization of COVID-19 Patient Sera ll Article Proteomic and Metabolomic Characterization of COVID-19 Patient Sera. 59–72 (2020). doi:10.1016/j.cell.2020.05.032
15. Bruzzone, C. *et al.* SARS-CoV-2 Infection Dysregulates the Metabolomic and Lipidomic Profiles of Serum. *iScience***23**, (2020).
16. Loo, R. L. *et al.* Quantitative In-Vitro Diagnostic NMR Spectroscopy for Lipoprotein and Metabolite Measurements in Plasma and Serum: Recommendations for Analytical Artifact Minimization with Special Reference to COVID-19/SARS-CoV-2 Samples. *J. Proteome Res.***19**, 4428–4441 (2020).
17. Barberis, E. *et al.* Large-Scale Plasma Analysis Revealed New Mechanisms and Molecules Associated with the Host Response to SARS-CoV-2. *Int. J. Mol. Sci.***21**, (2020).
18. Frampas, C. *et al.* Untargeted saliva metabolomics reveals COVID-19 severity. *medRxiv* (2021). doi:https://doi.org/10.1101/2021.07.06.21260080

19. Spick, M. *et al.* Changes to the sebum lipidome upon COVID-19 infection observed via rapid sampling from the skin. *EClinicalMedicine* 100786 (2021). doi:https://doi.org/10.1016/j.eclinm.2021.100786
20. Nascimento, M. H. C. *et al.* Noninvasive Diagnostic for COVID-19 from Saliva Biofluid via FTIR Spectroscopy and Multivariate Analysis. *Anal. Chem.* (2022). doi:10.1021/acs.analchem.1c04162
21. Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nat. Rev. Genet.***19**, 299–310 (2018).
22. Song, J. W. *et al.* Omics-Driven Systems Interrogation of Metabolic Dysregulation in COVID-19 Pathogenesis. *Cell Metab.***32**, 188-202.e5 (2020).
23. Wollina, U., Karadağ, A. S., Rowland-Payne, C., Chiriac, A. & Lotti, T. Cutaneous signs in COVID-19 patients: A review. *Dermatol. Ther.***33**, (2020).
24. Daneshgaran, G., Dubin, D. P. & Gould, D. J. Cutaneous Manifestations of COVID-19: An Evidence-Based Review. *Am. J. Clin. Dermatol.***21**, 627–639 (2020).
25. Knight, S. R. *et al.* Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. *BMJ***370**, m3339 (2020).
26. Prall, S. P. & Muehlenbein, M. P. DHEA Modulates Immune Function: A Review of Evidence. *Vitam. Horm.***108**, 125–144 (2018).
27. Butcher, S. K. *et al.* Raised cortisol: DHEAS ratios in the elderly after injury: Potential impact upon neutrophil function and immunity. *Aging Cell***4**, 319–324 (2005).
28. Trivedi, D. K. *et al.* Discovery of Volatile Biomarkers of Parkinson's Disease from Sebum. *ACS Cent. Sci.***5**, 599–606 (2019).
29. Sinclair, E. *et al.* Metabolomics of sebum reveals lipid dysregulation in Parkinson's disease. *Nat. Commun.***12**, (2021).
30. Rossini, E. L., Milani, M. I., Lima, L. S. & Pezza, H. R. Paper microfluidic device using carbon dots to detect glucose and lactate in saliva samples. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.***248**, 119285 (2021).
31. Do, K. T. *et al.* Network-Based Approach for Analyzing Intra- and Interfluid Metabolite Associations in Human Blood, Urine, and Saliva. *J. Proteome Res.***14**, 1183–1194 (2015).
32. Bel'skaya, L. V, Sarf, E. A. & Kosenok, V. K. Age and gender characteristics of the biochemical composition of saliva: Correlations with the composition of blood plasma. *J. Oral Biol. Craniofacial Res.***10**, 59–65 (2020).
33. Marksteiner, J., Oberacher, H. & Humpel, C. Acyl-Alkyl-Phosphatidlycholines are Decreased in Saliva of Patients with Alzheimer's Disease as Identified by Targeted Metabolomics. *J. Alzheimer's Dis.***68**, 583–589 (2019).
34. Nardo, A. D. *et al.* Pathophysiological mechanisms of liver injury in COVID-19. *Liver Int.***41**, 20–32 (2021).

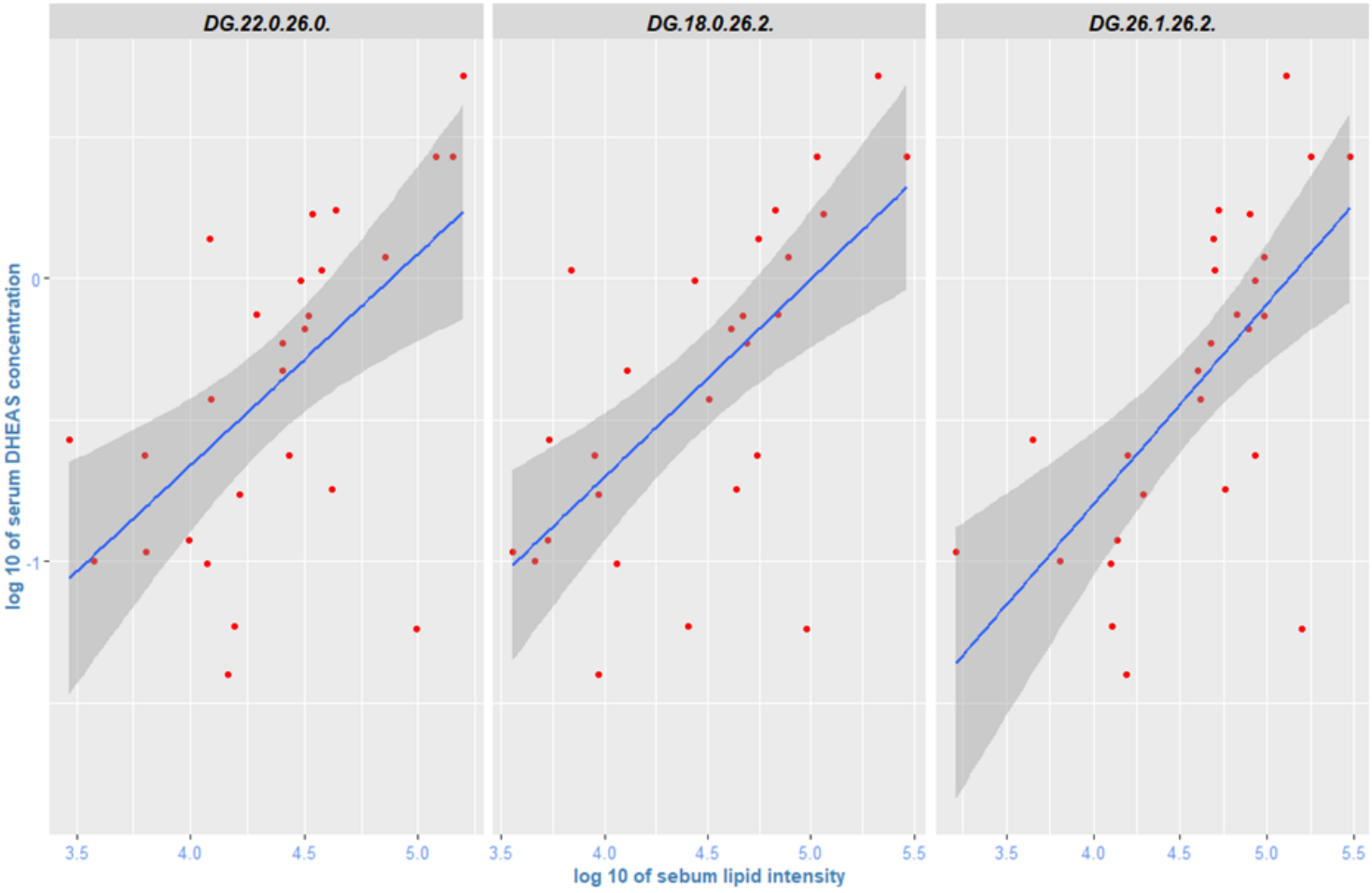
35. Castañé, H. *et al.* Machine learning identified distinct serum lipidomic signatures in hospitalized COVID-19-positive and COVID-19-negative patients. (2021). doi:https://doi.org/10.1101/2021.12.14.21267764
36. COVID-19 Mass Spectrometry Coalition. COVID-19 metabolomics and lipidomics protocol. (2020). Available at: https://covid19-msc.org/metabolomics-and-lipidomics-protocol/. (Accessed: 25th May 2021)
37. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics***3**, 211–221 (2007).
38. Chong, J., Wishart, D. S. & Xia, J. Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. *Curr. Protoc. Bioinforma.***68**, e86 (2019).
39. Python. (2021).
40. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.***12**, 2825–2830 (2011).
41. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.***40**, 16–28 (2014).
42. Struwe, W. *et al.* The COVID-19 MS Coalition—accelerating diagnostics, prognostics, and treatment. *The Lancet***395**, 1761–1762 (2020).

# Figures



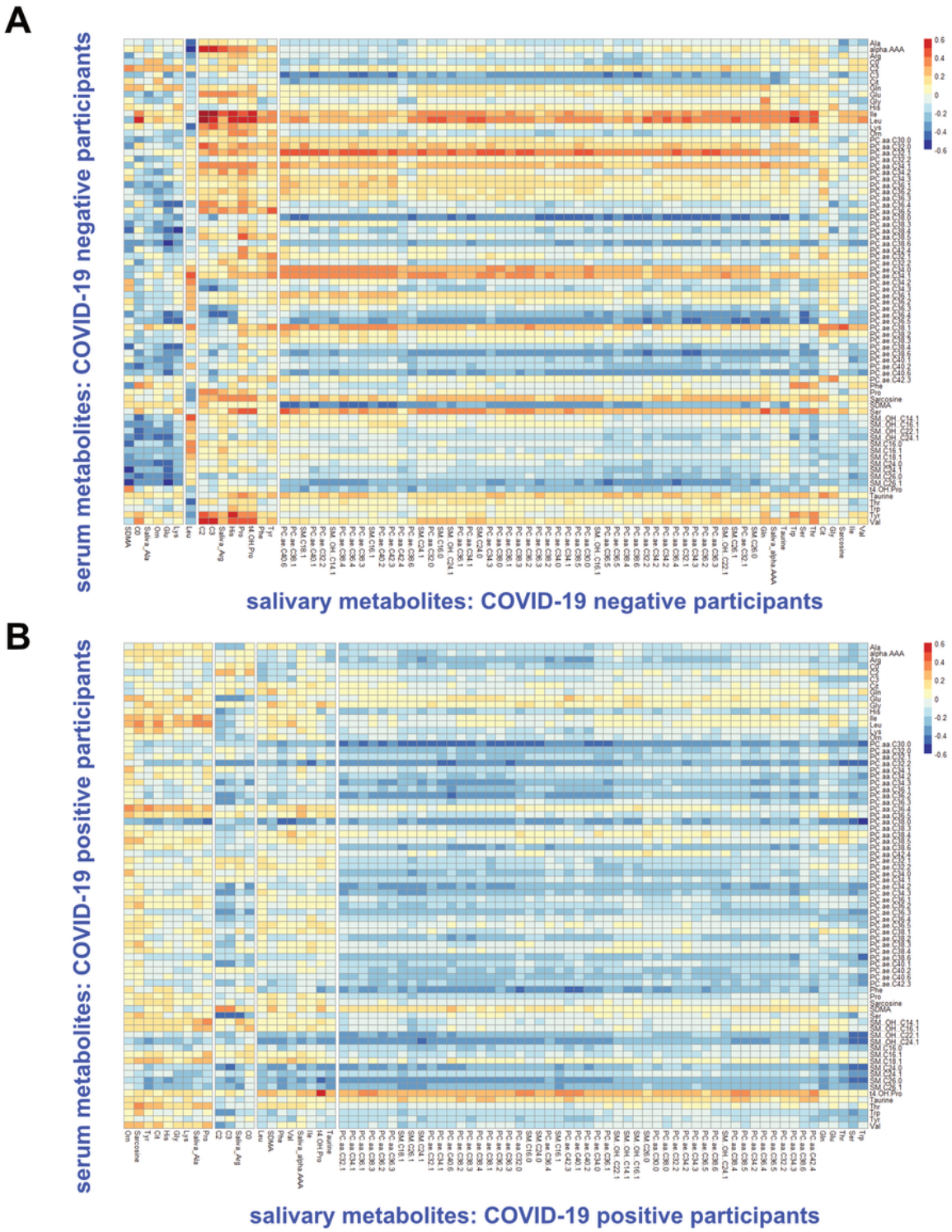

## Figure 1

Heatmaps showing Pearson correlation coefficients between sebum lipids and serum metabolites: COVID-19 negative participants (A) and COVID-19 positive participants (B)



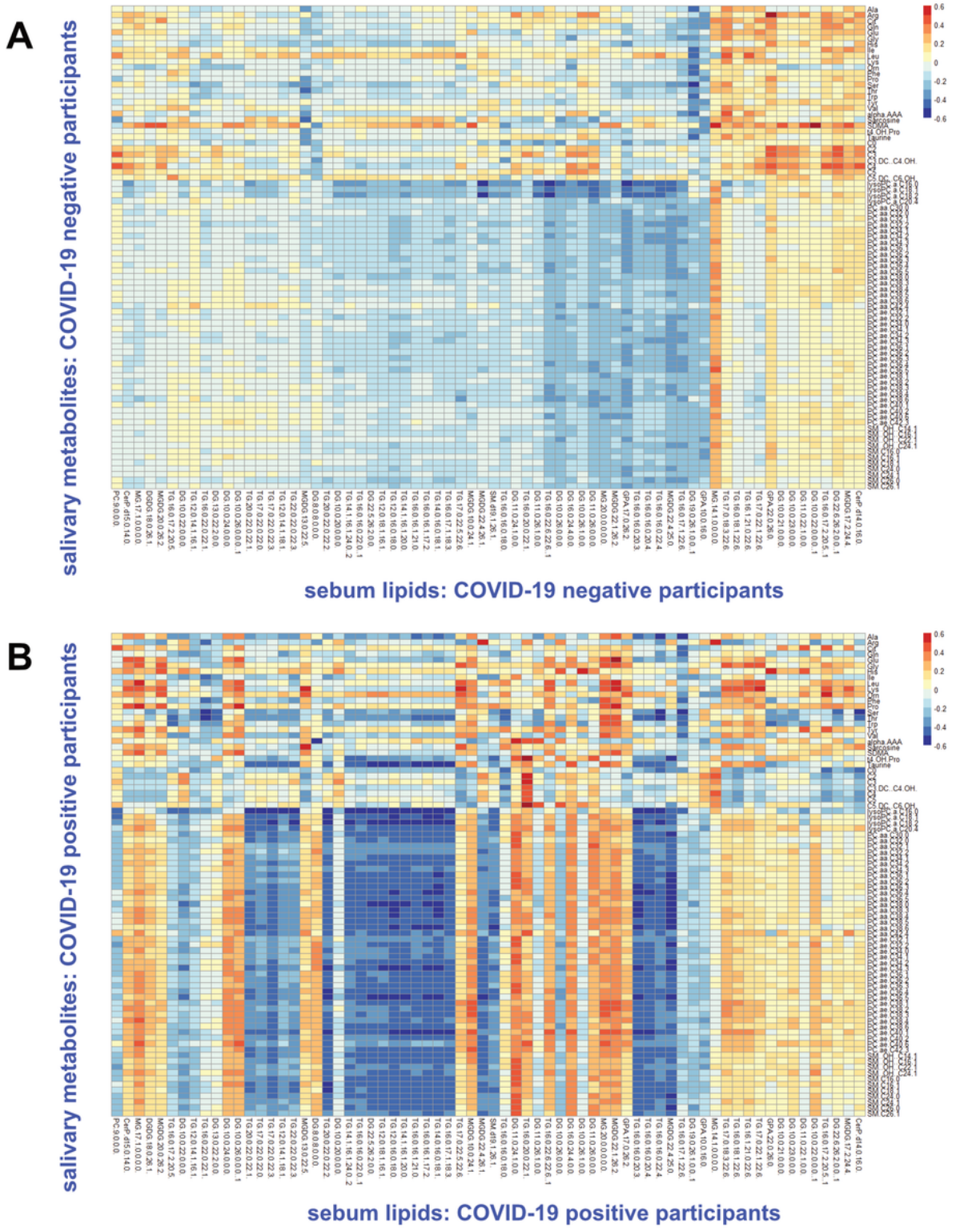

## Figure 2

Pairwise correlations between serum DHEAS and sebum diglycerides (COVID-19 Positive)



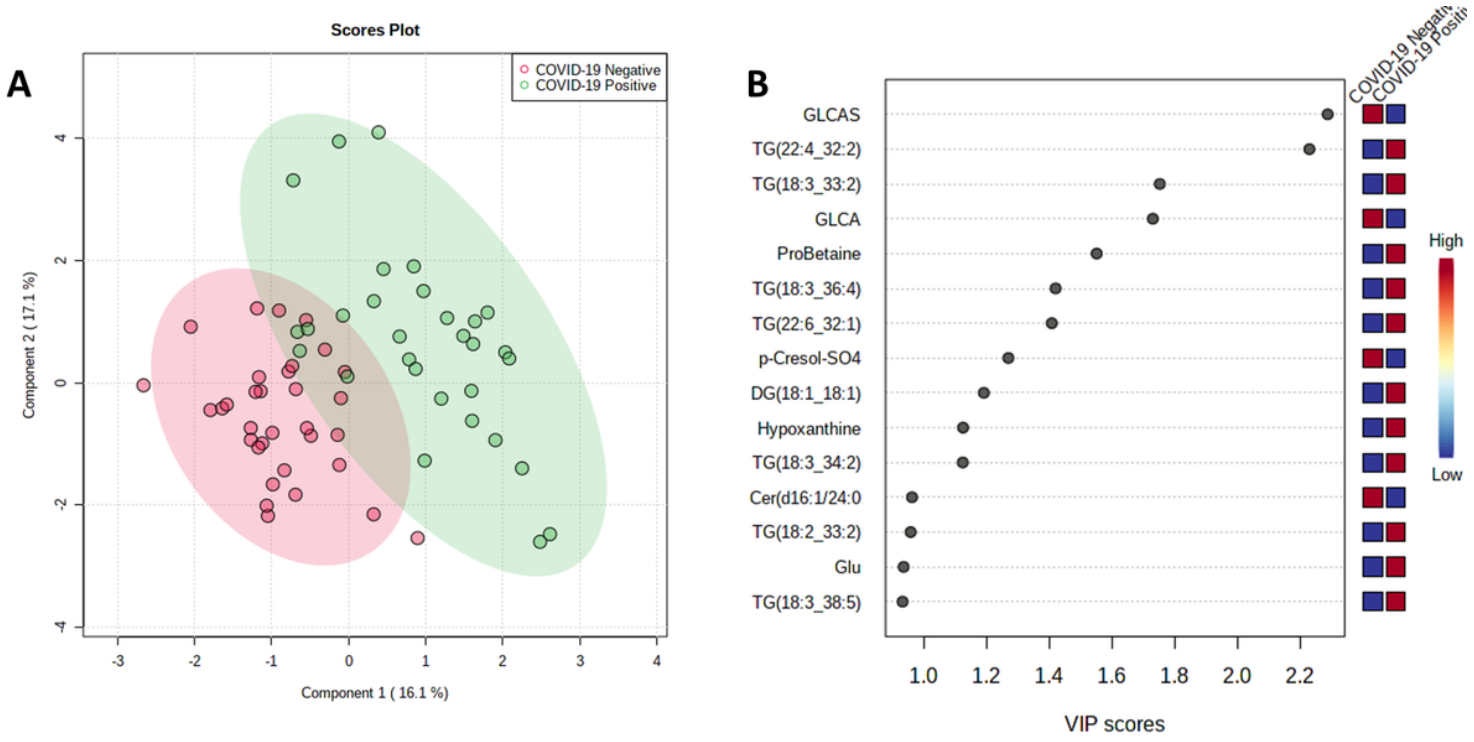

## Figure 3

Heatmaps showing Pearson correlation coefficients between serum and salivary metabolites: COVID-19 negative participants (A) and COVID-19 positive participants (B)



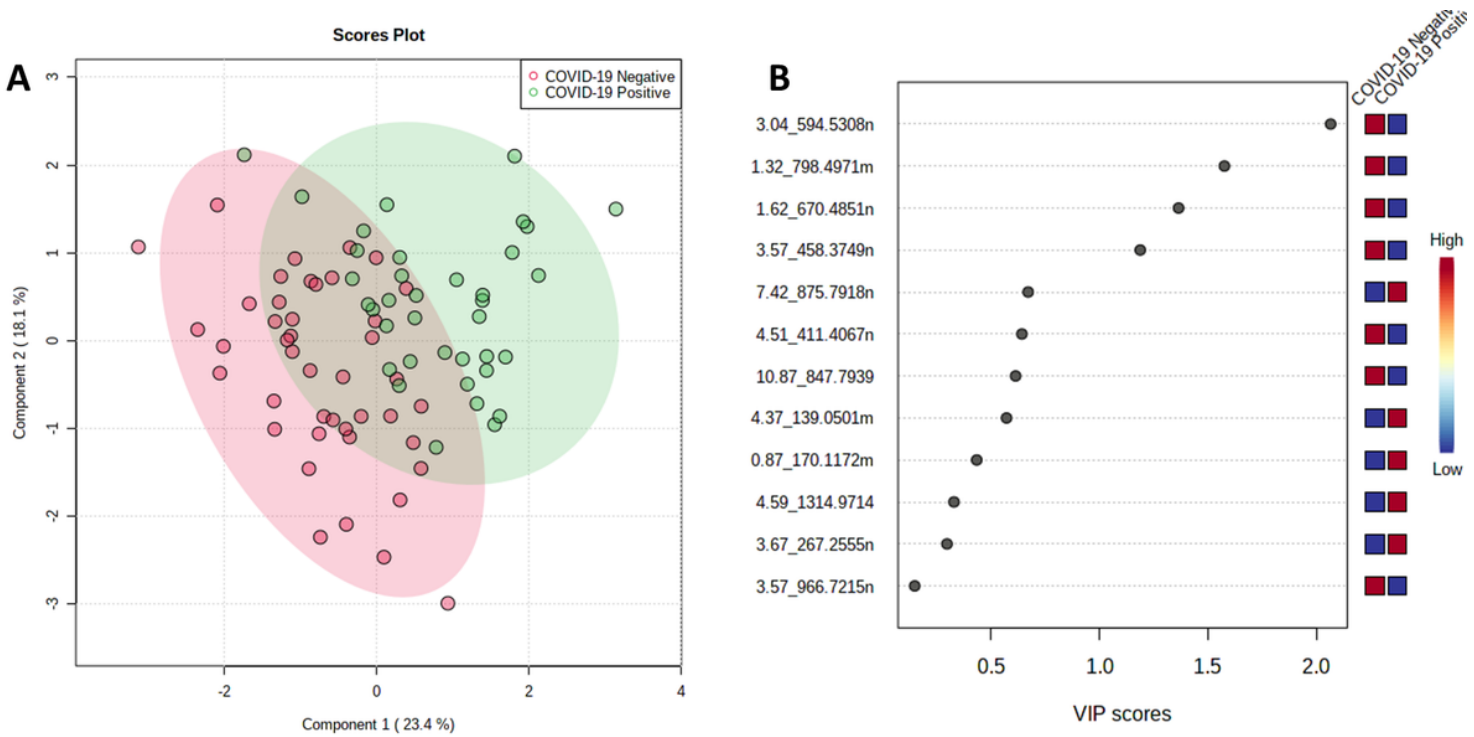

## Figure 4

Heatmaps showing Pearson correlation coefficients between serum and salivary metabolites: COVID-19 negative participants (A) and COVID-19 positive participants (B)




## Figure 5

(A) PLS-DA plot and (B) high VIP score metabolites for serum, COVID-19 positive versus COVID-19 negative




## Figure 6

(A) PLS-DA plot and (B) high VIP score metabolites for sebum, COVID-19 positive versus COVID-19 negative

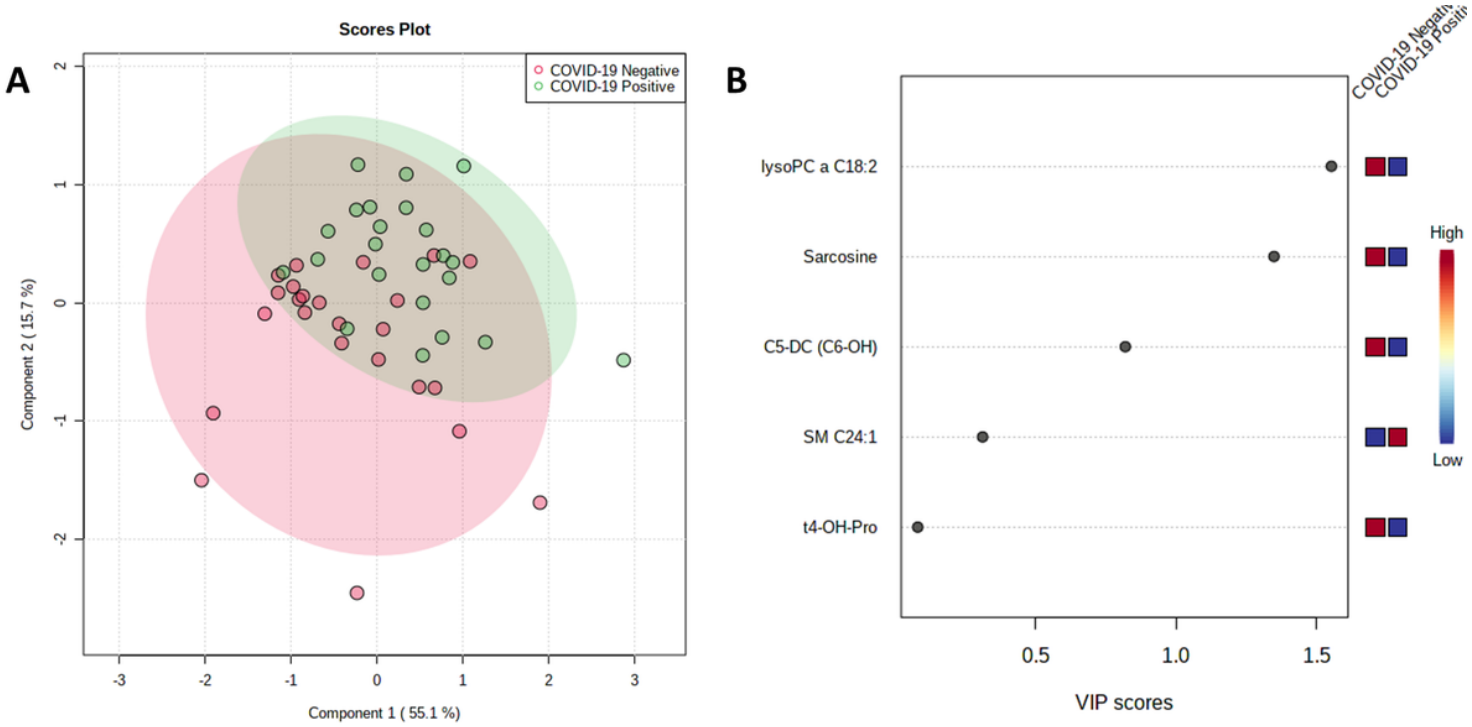



**Figure 7**

(A) PLS-DA plot and (B) high VIP score metabolites for saliva, COVID-19 positive versus COVID-19 negative

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BiofluidCorrelationsSuppMaterial16Feb2022.docx
- UntargetedMetabolomicsforDiscoveryofDiseaseBiomarkersGraphicalAbstract.png
- BiofluidCorrelationsSuppMaterial16Feb2022.docx
- FigureS1VennDiagram.png
- FigureS2sebumPCAs.png
- FigureS3serumPCAs.png
- FigureS4salivaPCAs.png