Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Gastro Hep Advances 2022;1:194–209

ORIGINAL RESEARCH—BASIC

A 'Multiomic' Approach of Saliva Metabolomics, Microbiota, and Serum Biomarkers to Assess the Need of Hospitalization in Coronavirus Disease 2019

Chiara Pozzi,[1,*] Riccardo Levi,[2,*] Daniele Braga,[1,*] Francesco Carli,[3] Abbass Darwich,[2] Ilaria Spadoni,[2] Bianca Oresta,[1] Carola Conca Dioguardi,[4] Clelia Peano,[4] Leonardo Ubaldi,[2] Giovanni Angelotti,[1] Barbara Bottazzi,[1] Cecilia Garlanda,[1,2] Antonio Desai,[1,2] Antonio Voza,[1,2] Elena Azzolini,[1,2] Maurizio Cecconi,[2,1] ICH COVID-19 Task-force[§,1,2], Alberto Mantovani,[1,2,5] Giuseppe Penna,[1] Riccardo Barbieri,[6] Letterio S. Politi,[1,2] and Maria Rescigno[1,2]

[1]IRCCS Humanitas Research Hospital, Rozzano, Milan, Italy, [2]Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy, [3]Department of Informatics, Università degli Studi di Torino, Torino, Piemonte, Italy, [4]Institute of Genetic and Biomedical Research, UoS of Milan, National Research Council, Rozzano, Milan, Italy, [5]The William Harvey Research Institute, Queen Mary University of London, London, UK, and [6]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy

BACKGROUND AND AIMS: The SARS-CoV-2 pandemic has overwhelmed the treatment capacity of the health care systems during the highest viral diffusion rate. Patients reaching the emergency department had to be either hospitalized (inpatients) or discharged (outpatients). Still, the decision was taken based on the individual assessment of the actual clinical condition, without specific biomarkers to predict future improvement or deterioration, and discharged patients often returned to the hospital for aggravation of their condition. Here, we have developed a new combined approach of omics to identify factors that could distinguish coronavirus disease 19 (COVID-19) inpatients from outpatients. **METHODS:** Saliva and blood samples were collected over the course of two observational cohort studies. By using machine learning approaches, we compared salivary metabolome of 50 COVID-19 patients with that of 270 healthy individuals having previously been exposed or not to SARS-CoV-2. We then correlated the salivary metabolites that allowed separating COVID-19 inpatients from outpatients with serum biomarkers and salivary microbiota taxa differentially represented in the two groups of patients. **RESULTS:** We identified nine salivary metabolites that allowed assessing the need of hospitalization. When combined with serum biomarkers, just two salivary metabolites (myo-inositol and 2-pyrrolidineacetic acid) and one serum protein, chitinase 3-like-1 (CHI3L1), were sufficient to separate inpatients from outpatients completely and correlated with modulated microbiota taxa. In particular, we found *Corynebacterium 1* to be overrepresented in inpatients, whereas *Actinomycetaceae F0332*, *Candidatus Saccharimonas*, and *Haemophilus* were all underrepresented in the hospitalized population. **CONCLUSION:** This is a proof of concept that a combined omic analysis can be used to stratify patients independently from COVID-19.

Keywords: Metabolome; Microbiota; CHI3L1; COVID-19

Introduction

The SARS-CoV-2 pandemic has drastically impacted on hospitals' beds and clinical practice. The choice of whether keeping a patient under treatment at hospital or discharge and treat them at home is at the discretion of the clinicians. It would be important for clinical and resource optimization purposes to predict who really needs

**These authors contributed equally to this work*

[§]ICH COVID-19 Task-Force: Aghemo Alessio, Anfray Clement, Badalamenti Salvatore, Belgiovine Cristina, Bertocchi Alice, Bombace Sara, Brescia Paola, Calcaterra Francesca, Calvi Michela, Cancellara Assunta, Capucetti Arianna, Carenza Claudia, Carloni Sara, Carnevale Silvia, Cazzetta Valentina, Cecconi Maurizio, Ciccarelli Michele, Coianiz Nicolò, Darwich Abbass, Lleo de Nalda Ana, De Paoli Federica, Di Donato Rachele, Digifico Elisabeth, Durante Barbara, Farina Floriana Maria, Ferrari Valentina, Fornasa Giulia, Franzese Sara, Gil Gomez Antonio, Giugliano Silvia, Gomes Ana Rita, Lizier Michela, Lo Cascio Antonino, Melacarne Alessia, Mozzarelli Alessandro, My Ilaria, Oresta Bianca, Pasqualini Fabio, Pastò Anna, Pelamatti Erica, Perucchini Chiara, Pozzi Chiara, Rimoldi Valeria, Rimoldi Monica, Scarpa Alice, Selmi Carlo, Silvestri Alessandra, Sironi Marina, Spadoni Ilaria, Spano' Salvatore, Spata Gianmarco, Supino Domenico, Tentorio Paolo, Ummarino Aldo, Valentino Sonia, Voza Antonio; Zaghi Elisa, Zanon Veronica.

Abbreviations used in this paper: **AUC**, area under the curve; **CHI3L1**, chitinase 3-like-1; **CI**, confidence interval; **COVID-19**, coronavirus disease 19; **DT**, decision tree; **ELISA**, enzyme-linked immunosorbent assay; **ESI**, electrospray ionization; **FDR**, false discovery rate; **IgG**, immunoglobulin G; **LR**, logistic regression; **PCA**, principal component analysis; **PTX3**, pentraxin 3; **RFE**, recursive feature elimination; **SVM**, support vector machine.

Most current article

2772-5723
https://doi.org/10.1016/j.gastha.2021.12.006

to remain at the hospital or be sent home without risks for the patient. SARS-CoV-2 transmission occurs primarily from liquid droplets entering contact with nasal and buccal mucosa (https://apps.who.int/iris/handle/10665/333114). Hence, these sites are the ones that most likely contribute to protecting from or facilitate the infection. SARS-CoV-2 enters the host via binding to the angiotensin-converting enzyme 2,[1] after proteolytic processing of the spike protein.[2] Many factors may contribute to infection at mucosal sites, including the metabolic output of the microbiota and the saliva.

During SARS-CoV-2 pandemic, we have learned that machine learning approaches may help stratify the patients and the use of multiomic approaches has proven useful to assess vaccine efficacy.[3] Differences in the plasma metabolome have been shown to predict coronavirus disease 19 (COVID-19) disease with an accuracy >74% and a sensitivity and specificity >75%, but not the clinical outcome.[4] Two pathways have been shown to be particularly discriminative: the tryptophan-nicotinamide pathway which is linked to inflammatory signals and to microbiota metabolism and the level of cytosine, shown to be pivotal in cell metabolism in the context of SARS-CoV-2 infection.[4] In another study, the serum lipidomics and metabolomics allowed to predict COVID-19 patients from healthy individuals, highlighting the impact that the infection has overall on individuals' metabolism.[5] In this case, an alteration of ketone bodies and redistribution of lipoprotein size were observed which are consistent with liver damage.[5] In addition, the perturbation of the metabolism of serum triglycerides and free fatty acids, especially arachidonic acid (area under the curve [AUC] = 0.99) and oleic acid (AUC = 0.98), correlated with the severity of the disease.[6] The serum anthranilic acid belonging to the kynurenine pathway also had a poor prognostic value in COVID-19 and was correlated with high interleukin-10 and -18 levels. Another study also clearly demonstrated that it was possible to assign 16 of 19 COVID-19 patients on the basis of plasma metabolome and proteomics.[7] They found a strong signature of innate immune cell dysregulation, including cytokine and complement system as well as a pronounced metabolomic suppression. Similarly, another study identified a mixture of 10 metabolites capable of distinguishing COVID-19 patients from the healthy population (AUC = 0.975). They found that COVID-19 plasma lipidomics resembled that of monosialodihexosyl ganglioside–enriched exosomes.[8] However, there have been no studies capable of distinguishing COVID-19 inpatients from outpatients.

Saliva metabolites can be the consequence of diet, host, or microbial metabolism. The salivary microbiota has been found to be altered in COVID-19 patients.[9] In particular, *Prevotella salivae* and *Veillonella infantium* were characteristic of the COVID-19 patients, whereas *Neisseria perflava* and *Granulicatella elegans* were predominant in controls. However, there are no data on the saliva metabolome and whether it correlates with the microbiome.

In this study, we analyzed the saliva metabolome and the serum of 50 COVID-19 patients (25 inpatients and 25 outpatients) and compared it with that of 270 healthy individuals having previously been exposed or not to SARS-CoV-2. We identified 9 metabolites that partly separated the populations of inpatients and outpatients. In addition, two of them (myo-inositol and 2-pyrrolidineacetic acid) when coupled to serum chitinase 3-like-1 (CHI3L1), an inflammatory protein shown to stimulate the expression of SARS-CoV-2 receptor angiotensin-converting enzyme 2 and to correlate with the severity of COVID-19,[10] allowed us to distinguish completely the two groups. We then correlated these 9 metabolites with differentially represented salivary microbiota taxa and identified the microbiota members positively or negatively associated with these metabolites.

Results

Metabolomic Profiles of COVID-19 Differently Exposed Individuals

We analyzed the saliva of 320 subjects, of which 50 were COVID-19 patients and 270 were either SARS-CoV-2–naïve healthy subjects (n = 180, immunoglobulin G [IgG] <12) or SARS-CoV-2–exposed individuals (IgG ≥12AU/mL, n = 90) who had been either asymptomatic/paucisymptomatic (n = 30) or symptomatic (n = 58) and who recovered from symptoms. Among the COVID-19 patients, 25 were hospitalized (inpatients) and 25 remained at home (outpatients). We collected their saliva and serum as closely as possible to SARS-CoV-2 nasal swab positivity. The characteristics of the COVID-19 patients are described in Tables 1 and 2.

First, we used a data set which comprised a total of 720 compounds of known identity (named biochemicals). After normalization to sample osmolality, log transformation, and imputation of missing values, if any, with the minimum observed value for each compound, Welch's two-sample *t*-test was used to identify biochemicals that differed significantly between experimental groups. A summary of the numbers of biochemicals that achieved statistical significance ($P \leq .05$), as well as those approaching significance ($.05 < P < .10$), is shown in Table A1.

We first carried out a principal component analysis (PCA) to obtain a high-level view of metabolomic data sets. With this approach, we reduced the dimensionality of the data while retaining most of the explained variance of the data set, allowing us to visually assess similarities and differences between samples. PCA of the global metabolite profiling data of saliva samples between all groups did not reveal clear separation of any specific group (Figure A1A) even when the COVID-19 patients were analyzed against the rest of the population (Figure A1B). This is likely due to that the variance associated with the first two components being 45.6%. We can assume that the metabolome of individuals with IgG positivity, as they had fully recovered from the disease (none of them was nasal swab positive at the time of

Table 1. Inpatients' Clinical Data

					Symptoms or clinical manifestations at hospitalization (yes = 1; no = 0)															WHO ordinal scale (from 0 to 8)			
Inpatients	Days between symptom onset and sample collection	Days between sample collection and hospitalization	Age	Sex	Fever	Low-grade fever	Headache	Cough	Sore throat	Muscle pain	Asthenia	Anosmia/ dysgeusia	Gastrointestinal symptoms	Conjunctivitis	Dyspnea	Chest pain	Tachycardia	Pneumonia	Others	Day 0	Day 1	Day 7	Day 13/ 14
3	12	2	56	M	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	3	3		
23	Around 10	3	75	F	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3	3		
15	60	1	42	F	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	2	3		
16	3	1	41	M	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	3	3		
6	Around 10	1	76	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3		
8	14	1	51	M	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	4	4	3	
1	18	3	41	M	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	3	3	3	
4	1	0	72	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	4	3	
19	33	3	76	M	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	3	3	3	
5	1	0	44	M	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	3	
13	Around 10	0	65	M	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	3	
12	3	1	73	M	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	3	
18	Around 10	2	69	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	3	3
17	24	13	85	M	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	3	3	3	3
21	4	3	89	F	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	3	3
22	Around 10	3	52	M	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	3	3	3	4
2	Less than 5	1	86	M	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	4	4	4	4
10	2	1	92	F	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	4	4	3	3
24	Around 10	1	100	F	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3	3	3	3
9	21	14	90	M	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	4	4	4	4
20	Less than 5	4	40	M	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	3	3
14	Less than 5	1	74	F	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	4	4	4	3

Table 1. Continued

Inpatients	Days between symptom onset and sample collection	Days between sample collection and hospitalization	Age	Sex	Symptoms or clinical manifestations at hospitalization (yes = 1; no = 0)												WHO ordinal scale (from 0 to 8)			
					Fever	Low-grade fever	Headache	Cough	Sore throat	Muscle pain	Asthenia	Anosmia/ dysgeusia	Gastrointestinal symptoms	Conjunctivitis	Dyspnea	Chest pain	Tachycardia	Pneumonia	Others	Day 0	Day 1	Day 7	Day 13/ 14
11	5	0	78	M	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	3	3	3	3
7	11	1	67	M	1	0	0	1	0	0	0	0	1	0	1	0	0	0	0	4	4	6	4
25	8	1	71	M	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	3	3	3	3

WHO ordinal scale: 0: uninfected; 1: no limitation of activities; 2: limitation of activities; 3: hospitalized, no oxygen therapy; 4: oxygen by a nasal mask; 5: noninvasive ventilation of high-flow oxygen; 6: intubation and mechanical ventilation; 7: ventilation + additional organ support; 8: death.
WHO, World Health Organization.

saliva testing), returned to a normal status. We then considered the possibility that the outpatients were somehow confounding the analysis of the COVID-19 population. Thus, we analyzed the population of COVID-19 inpatients vs the outpatients and the rest of the population. Even if analyzed separately, the three populations were still not distinguishable on a PCA (Figure A1C). Then, we analyzed only the groups of hospitalized vs nonhospitalized patients. As shown in Figure 1A, the two groups tended to separate, but there was no clear distinction of the two groups.

Nine Metabolites Distinguish the COVID-19 Population in Inpatients vs Outpatients

As the groups of COVID-19 inpatients and outpatients did not clearly separate, we considered the possibility that the amount of analyzed metabolites was diluting the information and hindering the separation. We thus used a machine learning approach to evaluate whether we could pinpoint some metabolites characterizing the two populations. As the hospitalized inpatients had an average age much higher than that of the other group (68.2 yo, std: 17.9 vs 41.4 yo, std: 9), to exclude age-related metabolite differences, we generated a model based on a training set of 80% of hospitalized patients (n = 20) and 5 patients age-matched randomly sampled from each class of the nonhospitalized patients to match the same number of hospitalized patients (n = 20). The other observations were inserted in the test set. We then performed an iterative bootstrap version based on the recursive feature elimination (RFE) algorithm,[11] where feature ranking could be evaluated by the number of times each feature was selected in a single iteration of the RFE algorithm. The first 9 features of the ranking were selected for further steps. These features clearly distinguished the populations of COVID-19 patients hospitalized (inpatients) from those who remained at home (outpatients). However, when we deeply analyzed the discriminating features, we found that 4 of 9 could be artifacts of dietary differences, possibly related to the hospitalization itself (caffeine-derived metabolites such as paraxanthine and theobromine) or to comorbidities, such as diabetes or obesity, such as sweeteners (acesulfame and erythritol) (Figure A2).

We thus proceeded in excluding possible confounding metabolites (n = 28) such as sweeteners, xanthine metabolites, and drugs (such as metformin) (Table A2) and reanalyzed the data.

After the exclusion of possible confounding metabolites, the algorithm was run again and 9 metabolites were identified as classifiers of the two groups (inpatients vs outpatients). These metabolites included the following: 2-pyrrolidineacetic acid (73 times), 1,3-diaminopropane (70 times), 3-hydroxypyridine (57 times), cyclo(leu-pro) (43 times), myo-inositol (38 times), N,N-dimethyl-5-aminovalerate (35 times), 3-(3-hydroxyphenyl)propionate (34 times), pantothenate (28 times), and mannonate* (25

Table 2. Outpatients' Clinical Data

				Symptoms or clinical manifestations (yes = 1; no = 0)															
Outpatients	Days between symptom onset and sample collection	Age	Sex	Fever	Low-grade fever	Headache	Cough	Sore throat	Muscle pain	Asthenia	Anosmia/ dysgeusia	Gastrointestinal symptoms	Conjunctivitis	Dyspnea	Chest pain	Tachycardia	Pneumonia	Others	WHO ordinal scale
30	3	50	F	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1
33	3	25	M	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1
35	3	28	F	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	1
36	25	41	F	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	2
37	2	33	M	1	0	0	0	1	1	0	1	1	0	0	0	0	0	0	1
43	34	30	F	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1
38	30	34	F	0	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1
39	50	58	M	1	0	1	0	1	1	1	1	0	1	0	0	0	0	0	1
40	31	43	F	0	1	0	1	1	0	1	1	0	0	0	0	0	0	0	1
41	38	53	M	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	2
42	2	44	F	1	0	1	1	1	1	1	1	0	0	1	0	1	0	0	2
44	3	41	F	1	0	1	0	1	0	0	1	0	0	0	0	0	0	0	1
45	19	38	F	1	0	1	0	1	1	1	1	1	0	0	1	0	0	0	1
46	23	53	M	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
47	3	41	F	1	0	0	1	1	1	1	1	0	0	1	1	1	0	0	2
48	23	37	F	1	0	0	1	1	1	0	0	0	0	1	0	0	0	0	1
26	2	39	F	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
27	1	56	M	1	0	1	0	1	1	1	0	0	0	0	0	0	0	0	1
49	37	50	F	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1
50	3	27	M	1	0	1	0	0	1	1	0	1	0	0	0	0	0	0	1
28	27	39	M	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
29	31	41	F	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	1
31	2	39	F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
32	30	48	F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
34	31	47	F	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	1

WHO ordinal scale: 0: uninfected; 1: no limitation of activities; 2: limitation of activities; 3: hospitalized, no oxygen therapy; 4: oxygen by a nasal mask; 5: noninvasive ventilation of high-flow oxygen; 6: intubation and mechanical ventilation; 7: ventilation + additional organ support; 8: death.
WHO, World Health Organization.

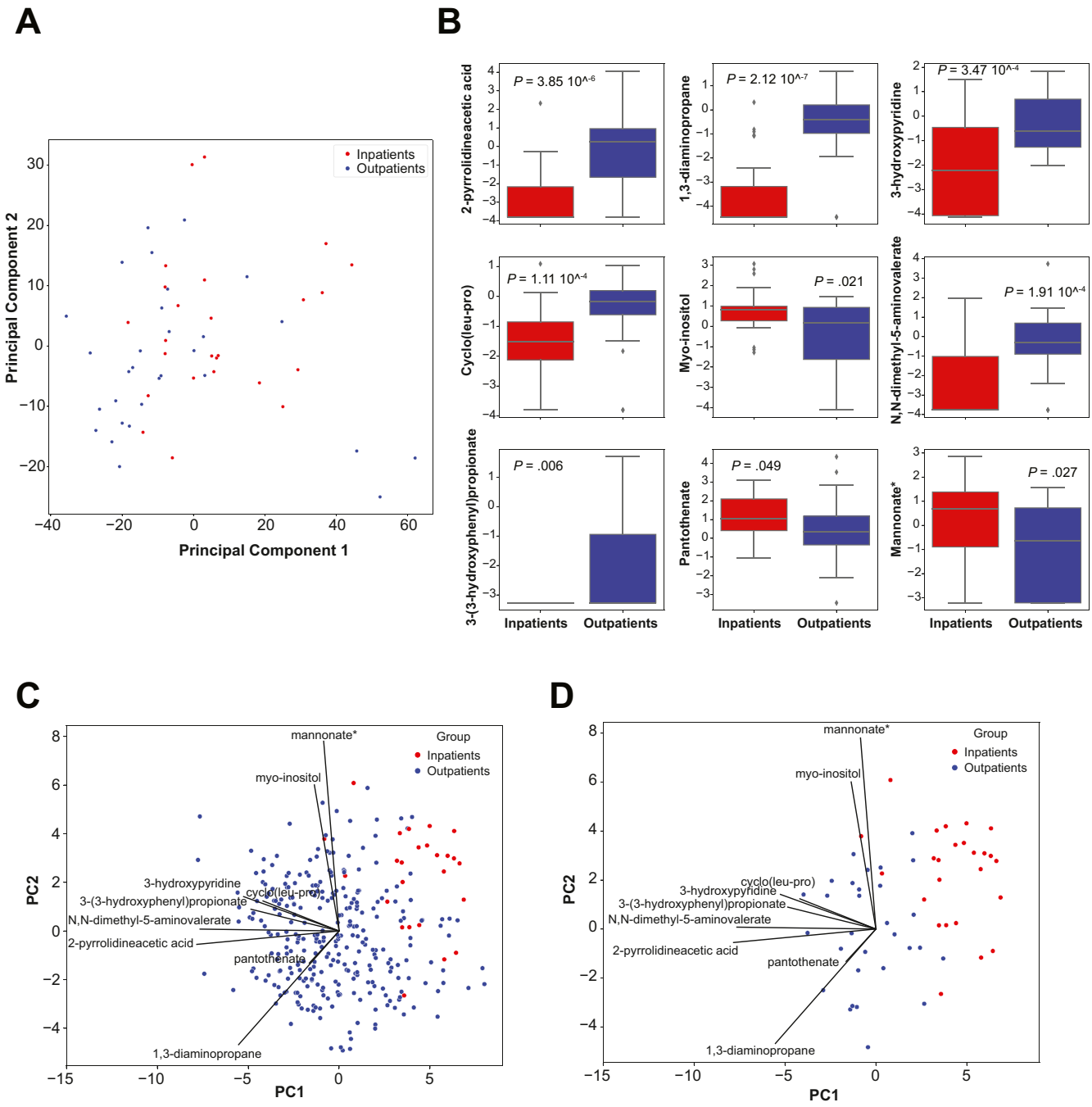

Figure 1. Statistical analysis of saliva metabolites characterizing the COVID-19 inpatients vs outpatients, excluding biased metabolites. (A) PCA of saliva metabolites. The plot shows PC1 and PC2 values of inpatients (n = 25) and outpatients (n = 25). (B) Boxplots showing the 9 metabolites detected from the feature selection algorithm in inpatients vs outpatients. Box plots show the interquartile range (IQR), the vertical lines show the median values, and the whiskers extend from the hinge no further than 1.5*IQR. (C) The PCA biplot of the best 9 metabolites in COVID-19 inpatients and the rest of the population. The explainable contribution of each metabolite in determining the direction of variance in the PCA space is represented by a solid line. (D) The PCA biplot of the best 9 metabolites in COVID-19 inpatients and outpatients. The explainable contribution of each metabolite in determining the direction of variance in the PCA space is represented by a solid line.

times) (Figure 1B and Table A3). The PCA considering these 9 metabolites improved the separation of the population of hospitalized COVID-19 patients from the rest of the population (Figure 1C) and from the outpatients but not completely (Figure 1D and Figure A3), and there were still some subjects who fell in the wrong group, and clinically, there was no reason to be considered as inappropriately being hospitalized or not.

CHI3L1, 2-Pyrrolidineacetic acid, and Myo-inositol Separate COVID-19 Inpatients From Outpatients

As we found that the nine metabolites could not completely separate the COVID-19 inpatients from the outpatients, we evaluated whether the combination of the 9 metabolites with serum features could help further separating the two groups. We tested 2 serum features: CHI3L1 and pentraxin 3 (PTX3), which were chosen as they are prognostic markers of severe disease,[10,12] and we found them to be statistically significantly higher in the inpatient population (Figure 2A–B) than in the outpatients. In addition, a Pearson correlation matrix clearly showed that the two serum features (PTX3 and CHI3L1) were correlated and several metabolites were equally correlated with each other (Figure 2C). We then applied a decision tree (DT) learning approach to assess if there were features that helped in discriminating the two populations and found that CHI3L1, but not PTX3, could separate the population into hospitalized individuals (CHI3L1 values $>$69.3 ng/ml) and outpatients (CHI3L1 values $\leq$69.3 ng/ml). As CHI3L1 was shown to correlate with age and thyroid cancer,[13] we assessed whether its serum concentrations were linked to age, but we can exclude that differences were simply related to age (Figure A4). However, the separation based on CHI3L1 was incomplete as 2 patients per group were not correctly assigned (Figure 2D). Correct separation of the two groups was achieved only with the help of the metabolites and, in particular, of two of the 9 metabolites identified previously (2-pyrrolidineacetic acid and myo-inositol). A value higher than the 0.442 scaled intensity value (log transformed) of 2-pyrrolidineacetic acid characterized the hospitalized population, whereas a value which was lower or equal to the 1.492 scaled intensity value (log transformed) of myo-inositol characterized the outpatients.

To evaluate whether this decision analysis was corroborated by a statistical learning framework, we reanalyzed the different variables with a support vector machine (SVM) approach. As shown in Figure 2E–F, the data were strongly supported by the SVM with 11 support vectors.

Hence, the combination of one serum biomarker (CHI3L1) and two metabolites (2-pyrrolidineacetic acid and myo-inositol) was sufficient to correctly assign the two populations with an accuracy of 86.4 (90% confidence interval [CI] 80.0–100.0) and area under the receiver characteristic operating curve of 95.2 (90% CI 91.9–100.0), evaluated in a 10-fold random permutated cross-validation.

The Salivary Microbiota Differs Between COVID-19 Inpatients and Outpatients

The 9 metabolites that we identified in the analysis can have several origins. They may derive from the diet, the microbiota, and mammalian cell metabolism. We thus first analyzed whether inpatients and outpatients had different salivary microbiota composition as measured by 16S rRNA analysis. We found that the microbiota in the two patient populations (COVID-19 inpatients vs outpatients) was greatly different. Both the alpha-diversity (evaluated as Shannon and Chao index) and beta-diversity (principal coordinates analysis) were statistically significantly different between inpatients and outpatients (Figure 3A–B). When analyzed at the genus level, there were clear differences between the two groups as shown by the volcano plot (Figure 3C and Figure A5). Forty-eight Operational Taxonomic Units (OTUs) differentially characterized the two populations, and these, except for 3 genera (*Corynebacterium 1*, *Rickettsiales mitochondria*, *Lactobacillus*), were mostly downregulated in the hospitalized population as compared with the outpatients. In Figure 3D and Figure A6, the differentially represented genera in the two groups are reported.

In the beginning of the pandemic, almost all the patients were treated with antibiotics which are known to strongly affect the microbiota.[14,15] Indeed, we found that—except for 5—all the inpatients had been treated with antibiotics for at least one day. Thus, we evaluated whether there were specific OTUs which were modified between hospitalized inpatients and outpatients and that were not affected by antibiotics. As shown in Figure 4A, the group of 5 patients not treated with antibiotics still had a microbiota composition which differed greatly from that of the outpatients ($P = .01$) and partly also with that of antibiotic-treated patients ($P = .044$). Eleven taxa were differentially represented in hospitalized patients not treated with antibiotics vs outpatients (Figure 4B). Of these 11 taxa, 4 had a similar pattern also in the other inpatients treated with antibiotics, thus indicating that these taxa were modified independently from antibiotic treatment in all our COVID-19 hospitalized inpatients. Among these strains, we found *Corynebacterium 1* to be overrepresented in the hospitalized population, whereas *Actinomycetaceae F0332*, *Candidatus Saccharimonas*, and *Haemophilus* were all underrepresented in the inpatient population (Figure 4C). *Candidatus Saccharimonas* is an obligate epibiont of the *Actinomyces odontolyticus* strain XH001.[16] We found that *A odontolyticus* strain XH001 was not affected in the two patient populations, confirming that XH001 does not depend on *Candidatus Saccharimonas* for its growth (Figure A7). These results show that the COVID-19 inpatients and outpatients have a different salivary microbiota composition.

2-Pyrrolidineacetic Acid and Myo-inositol Correlate With Specific Microbiota Genera

Finally, we analyzed whether there was a correlation with the differentially represented taxa of the salivary microbiota and the 9 metabolites differentiating the two populations. We thus carried out each pairwise combination of microbial genus relative abundances and the 9 metabolite intensities with at least one significant correlation (false discovery rate [FDR] $<$0.05) for downstream analysis and representation. The correlation coefficients of the 9 selected metabolites vs the microbial abundances were then

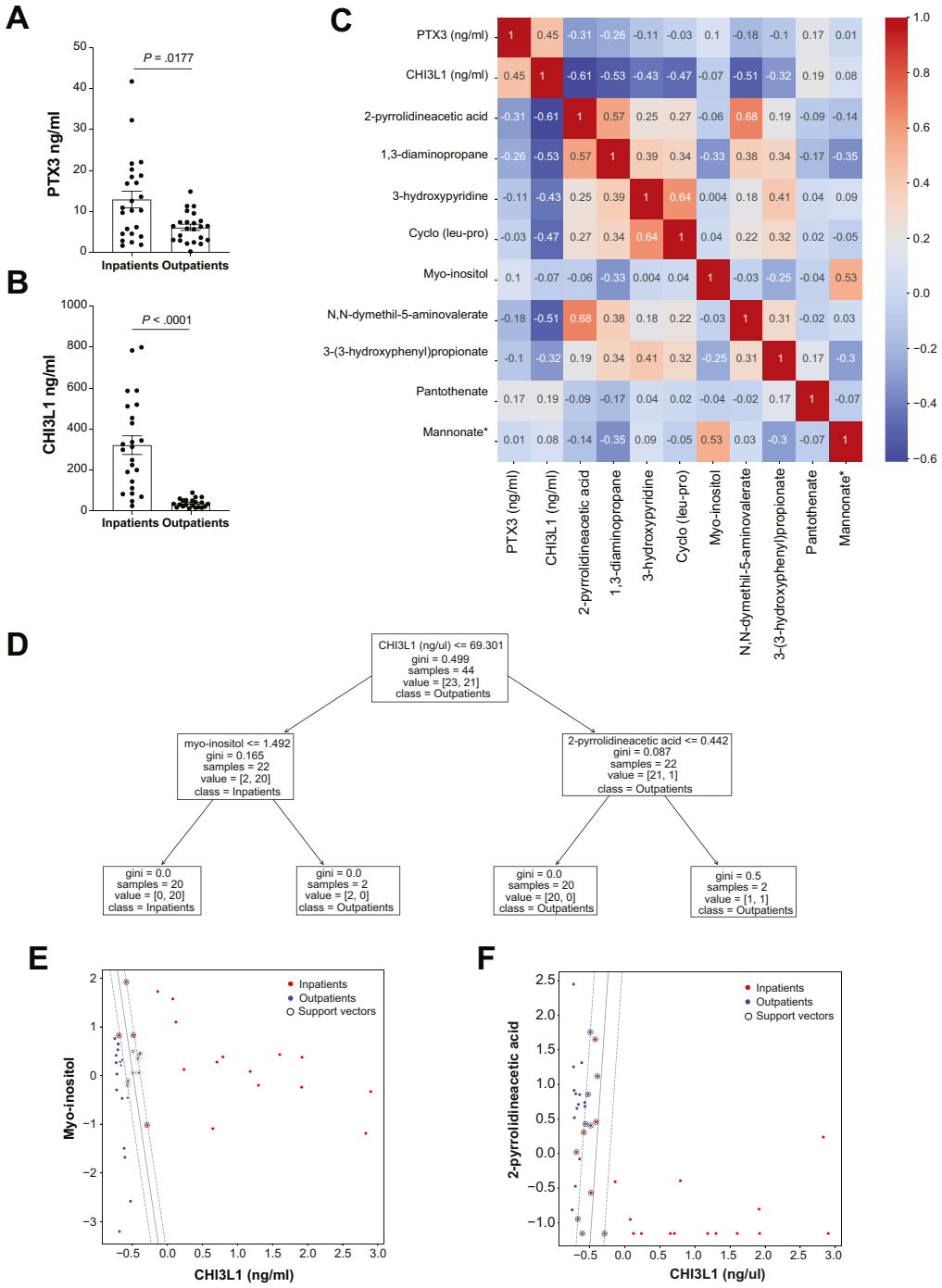

Figure 2. 2-pyrrolidineacetic acid and myo-inositol metabolites and CHI3L1 distinguish COVID-19 hospitalized patients vs outpatients. (A) CHI3L1 levels measured in the plasma of COVID-19 inpatients (n = 24) and outpatients (n = 21). Data are represented as the mean ± s.e.m. A two-tailed unpaired t-test was performed. (B) PTX3 levels measured in the plasma of COVID-19 inpatients (n = 24) and outpatients (n = 23). Data are represented as the mean ± s.e.m. A two-tailed Mann-Whitney test was performed. (C) The correlation matrix between the best 9 saliva metabolites and PTX3 and CHI3L1 plasma levels. In each cell, the Spearman correlation coefficient is reported. (D) The outline of the decision tree for the discrimination between COVID-19 inpatients vs outpatients. In each node, the splitting rule, the Gini value, the number of samples, and the relative distribution in each class are reported. (E) Boundary analysis with SVMs between CHI3L1 and myo-inositol. The straight line represents the decision boundary; the dotted lines are the margins, and black circles are the support vectors. (F) Boundary analysis with SVMs between CHI3L1 and 2-pyrrolidineacetic acid. The straight line represents the decision boundary; the dotted lines are the margins, and black circles are the support vectors.

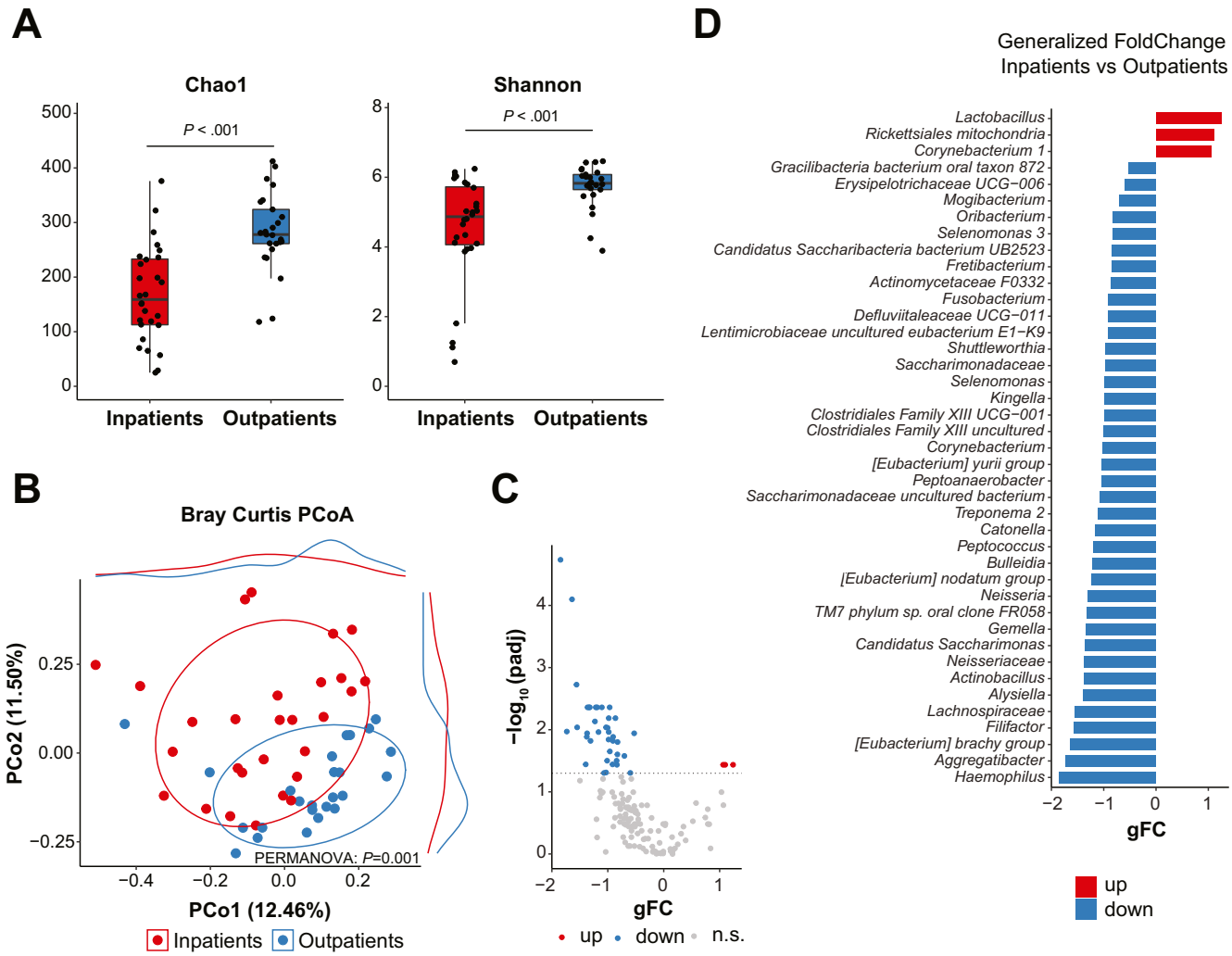

Figure 3. The microbiota changes in COVID-19 hospitalized patients vs outpatients. (A) Chao1 and Shannon alpha-diversity metrics of outpatients (n = 25) and inpatients (n = 28). Box plots show the interquartile range (IQR), the vertical lines show the median values, and the whiskers extend from the hinge no further than 1.5*IQR. (B) PCoA of inpatients and outpatients as per the Bray-Curtis distance matrix. (C) The volcano plot of differentially abundant bacteria between inpatients and outpatients. (D) Generalized fold change of bacteria showing significant different abundance in inpatients and outpatients. PCoA, principal coordinates analysis.

represented by a heatmap with hierarchical clustering. As shown in Figure 5A, as per metabolite correlations, the genera could be separated into 3 clusters. Interestingly, the 4 antibiotic-independent genera that we found to be either downregulated (*Haemophilus*, *Actinomycetaceae* F332, *Candidatus Saccharimonas*) or upregulated (*Corynebacterium 1*) in COVID-19 inpatients vs outpatients belonged, respectively, to cluster 1 or cluster 3. We then confirmed this finding in a spearman correlation analysis between the 4 genera and the 9 metabolites or CHI3L1. We found that each one of the taxa was associated positively or negatively with the 9 metabolites, but one bacterium in particular (*Candidatus Saccharimonas*) correlated with both myo-inositol and 2-pyrrolidineacetic acid in an opposite way (Figure 5B). It was inversely correlated with myo-inositol and positively with 2-pyrrolidineacetic acid. Interestingly *Candidatus Saccharimonas* was also negatively correlated with CHI3L1, whereas the latter was positively correlated with *Corynebacterium 1* (Figure 5C). These results indicate that there is a correlation between the salivary microbiota and the metabolites that characterize the populations of COVID-19 inpatients and outpatients.

Conclusion

In this study, we applied a multiomic approach to identify markers that are able to distinguish the inpatient from the outpatient COVID-19 population. We focused on the saliva, a site where the virus is detected for a long time and can be infected via the oral epithelium,[17-19] and analyzed both the metabolome and microbiota. Changes in the composition of the gut microbiota[20-22] and mycobiota,[23] as well as nasopharyngeal[24,25] and lower airway microbiota,[26] have been

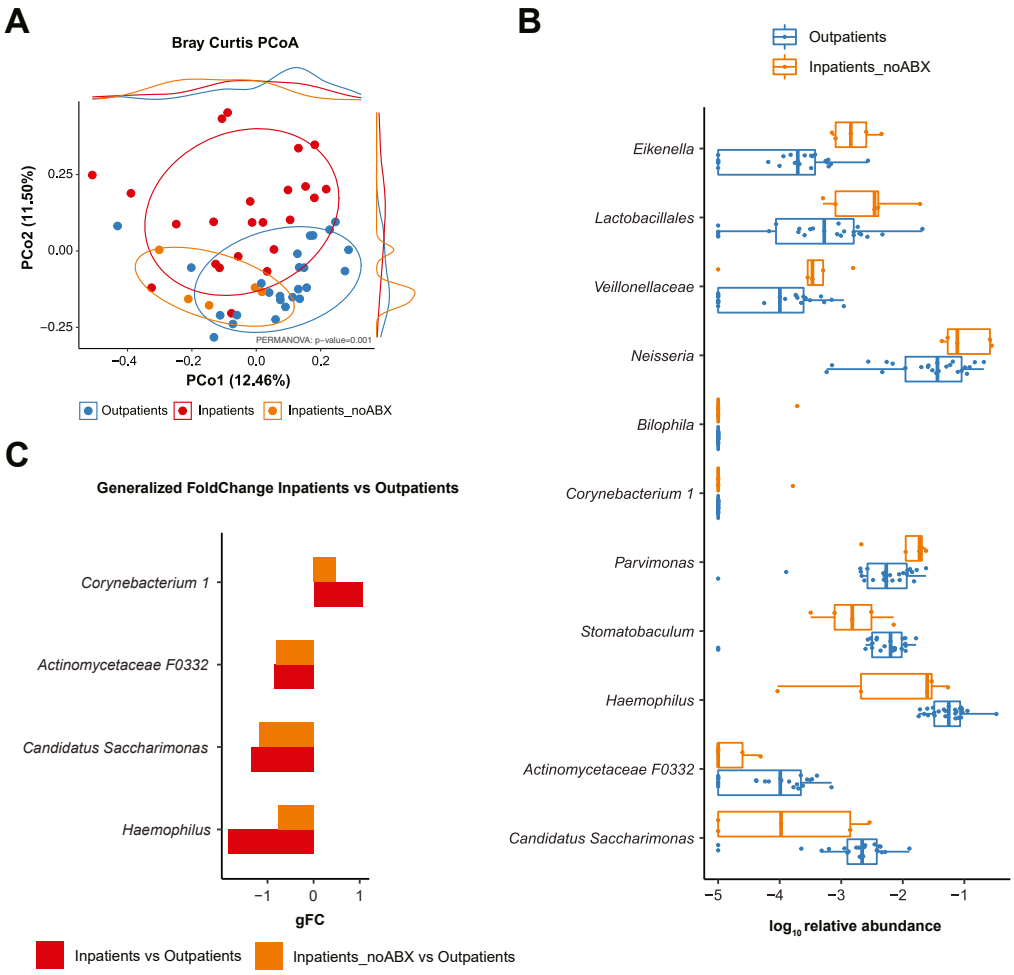

Figure 4. Identification of bacteria that are not influenced by antibiotics. (A) PCoA of COVID-19 inpatients treated with antibiotics (n = 23) and outpatients (n = 25) as per the Bray-Curtis distance matrix. Inpatients who did not receive antibiotic therapy are highlighted in orange (n = 5). (B) Boxplots of relative abundance (log_{10} scale) of bacteria with significant different abundance in inpatients not receiving antibiotics and outpatients. (C) Generalized fold change of bacteria showing significant different abundance in inpatients not receiving antibiotics and outpatients. Box plots show the interquartile range (IQR), the vertical lines show the median values, and the whiskers extend from the hinge no further than 1.5*IQR. PCoA, principal coordinates analysis.

observed in COVID-19 patients, with similar restoration times in both the nasal and gut microbiota composition.[27] In addition, a change in the microbiota correlated with a change in fecal metabolome,[21] and the microbiota has been shown to control the host glycosaminoglycan heparan sulfate through the activity of glycosidases, and this correlated with viral adhesion and infectivity.[28] Hence, both the microbiota and its metabolome may participate to increase the risk of SARS-CoV-2 infection and its subsequent clinical outcome. During the peak of the pandemic, the hospitals were tremendously under pressure and the decision to whether hospitalize the patients or not was at the discretion of the clinicians. However, in many cases, it was very difficult to predict what would be the outcome for an individual patient. Here, we evaluated whether there were metabolites characteristic of either one or the other population. We applied a machine learning approach to identify those metabolites which better separated the 2 COVID populations. In our first attempt, we found that 4 of the 9 metabolites best differentiating the two populations comprised several diet-related metabolites, including caffeine-derived metabolites (such as paraxanthine and theobromine) more abundant in outpatients and sweeteners (acesulfame and erythritol) more abundant in inpatients. As the hospitalized patients were not exposed to caffeine and tended to use sweeteners that could have been associated with their comorbidities (obesity, diabetes), we decided to repeat the analysis and exclude these and other diet-related metabolites. In the second round, we identified another set of metabolites based on how many times they were selected in a single iteration of the RFE algorithm: 2-pyrrolidineacetic acid, 1,3-diaminopropane, 3-hydroxypyridine, cyclo(leu-pro), myo-inositol, N,N-dimethyl-5-aminovalerate, 3-(3-hydroxyphenyl)propionate, pantothenate, and mannonate*. The pathways leading to the generation of these metabolites are known except for 2-pyrrolidineacetic acid, 3-hydroxypyridine, and cyclo(leu-pro) which are xenobiotics. However, the pyrrolidine ring of 2-pyrrolidineacetic acid can be linked to polyamine and, in particular, putrescine catabolism (Figure A8). 2-pyrrolidineacetic and 1,3-diaminopropane (also generated by polyamine catabolism) were found to be significantly more abundant in the outpatient population (Table A3) and to positively correlate with 2 taxa (*Haemophilus* and *Candidatus Saccharimonas*) all downregulated in the inpatient population. 1,3-diaminopropane was also negatively correlated with *Corynebacterium 1* which was highly represented in the inpatient population. These results indicate that the polyamine metabolism may be affected during COVID-19

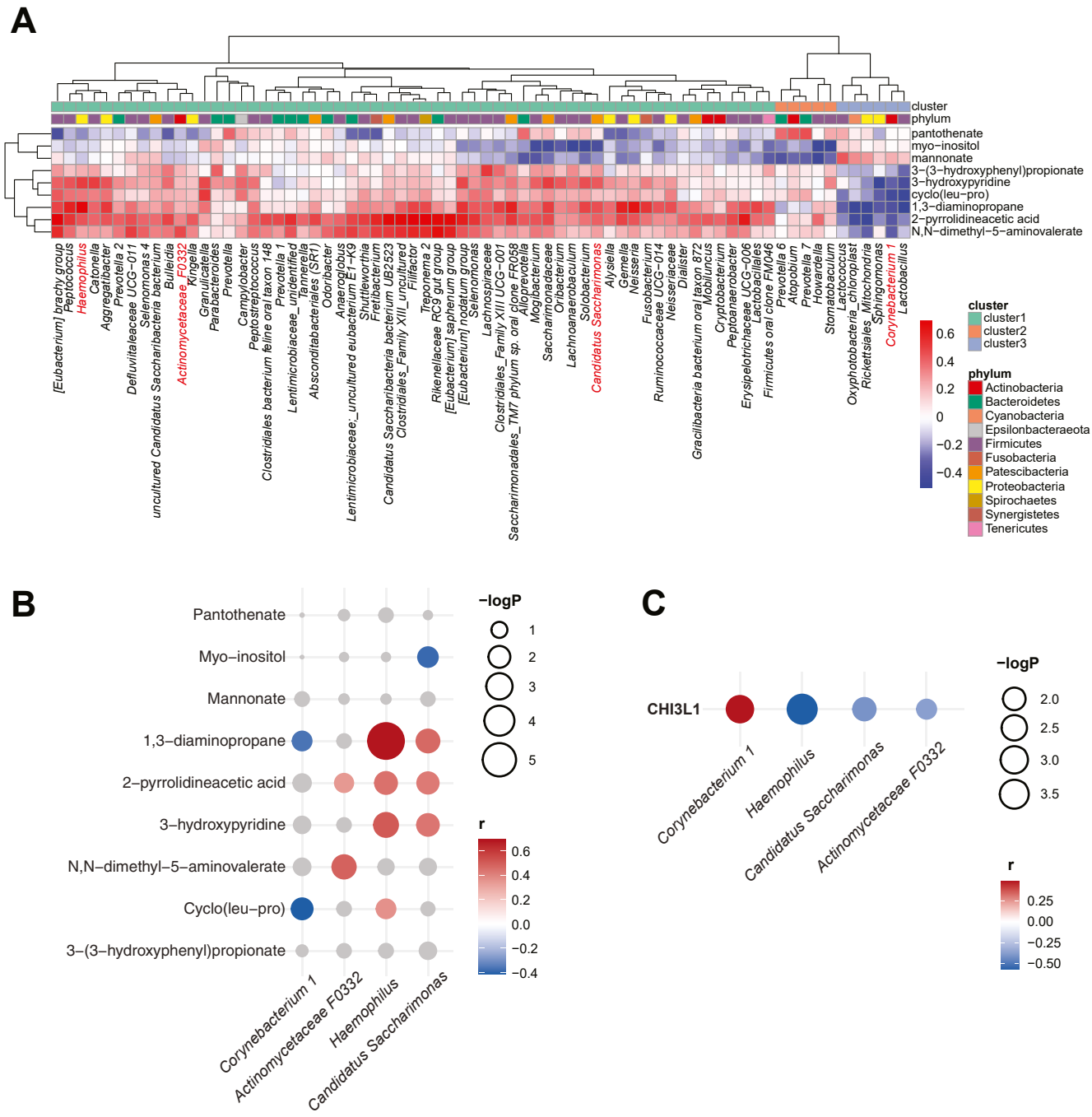

Figure 5. Spearman correlation of the microbial genus relative abundances vs metabolite intensities and serum CHI3L1. (A) The heatmap with hierarchical clustering representing the correlation coefficients of the 9 selected metabolites vs the microbial genus abundances. Columns refer to the microbial genera and were provided with two colored annotation bars, respectively, highlighting the taxonomic classification at the phylum level and the clusterization obtained with the dendrogram. (B) The bubble plot showing the results of the Spearman correlation analysis of the abundance of 4 selected bacteria vs the abundance of 9 selected metabolic features. The size of the bubble and the colored scale, respectively, represent the level of significance of the correlation (false discovery rate) and the correlation coefficient. The four bacteria were identified as being not affected by antibiotics in hospitalized patients and being differentially represented in hospitalized vs outpatients. (C) The bubble plot showing the results of Spearman correlation analysis of the abundance of 4 selected bacteria vs serum CHI3L1.

progression. Polyamines, and in particular putrescine, have been shown to facilitate coronavirus replication, attachment, and entry into cells,[29] whereas 1,3-diaminopropane interferes with the replication of the Semiki Forest virus (a highly replicative RNA virus).[30] Thus, the finding that 1,3-diaminopropane together with other putrescine-derived metabolites is higher in outpatients and is discriminative of the inpatient and outpatient populations may suggest that

these metabolites contrast viral replication or attachment, presumably via the degradation of putrescine. Their reduction may render the host more susceptible to viral infection.

Candidatus Saccharimonas, previously called TM7, is found in low abundance in healthy individuals.[31,32] It has an ultrasmall size (between 200 and 300 nm)[16] and is an obligate epibiont of the *A odontolyticus* strain XH001.[16] Interestingly, TM7 controls the growth of XH001, suggesting a parasitic relationship,[33] but it also gives a survival advantage to XH001 by inhibiting Tumor Necrosis Factor-Alpha (TNF-α) expression in macrophages, thus affecting its detection and elimination by phagocytes.[16] It is likely that the absence of TM7 has positive and negative effects on XH001, and this results in no changes in XH001 abundance, consistent with our observation. However, as TM7 controls the production of TNF-α by macrophages, its absence may favor the production of TNF-α in inpatients, which could contribute to increased inflammation.

By applying a DT and an SVM approach, we found that 2-pyrrolidineacetic acid with myo-inositol and serum CHI3L1 could completely distinguish the inpatient from the outpatient population. Interestingly, previously, it was shown that serum amounts of CHI3L1, age, and other factors were predictive of more severe COVID-19.[10] In our cohort, we show that CHI3L1 levels were elevated in hospitalized patients compared with age-matched healthy individuals and that together with the two abovementioned metabolites could also distinguish outpatients from inpatients with nonfatal COVID-19. 2-pyrrolidineacetic acid can be acquired from plants (in particular Tussilago farfara, tobacco leaves, and flower heads of several *Arnica* spp, such as *montana, chamissonis, amplexicaulis*, and *sachalinensis*, and green tea leaves), whereas the pyrrolidine ring, as mentioned previously, may also derive from the catabolism of putrescine. Hence, it is difficult to identify the exact source of 2-pyrrolidineacetic acid. Its precise activity is also unknown, and we cannot exclude that deriving from putrescine catabolism its presence may rather signify the elimination of a molecule involved in viral entry.

Myo-inositol is produced from glucose in eukaryotic cells and in some prokaryotic cells,[34] but it is not a sugar and can be acquired from several foods (particularly from grains). Myo-inositol can be used as a carbon source by several microorganisms,[35] and we found it to be negatively correlated with *Candidatus Saccharimonas*. Interestingly, myo-inositol is administered to preterm babies at risk of respiratory distress syndrome[36] for its capacity to improve lung surfactant properties.[37] Hence, it is not surprising that it is elevated in inpatients who have a more severe disease, but none of them succumbed from the COVID-19. In addition, TNF-α significantly reduces intracellular myo-inositol accumulation and its metabolism, resulting in the alteration of endothelial cellular function.[38] It may be possible that the reduction of *Candidatus Saccharimonas* may favor TNF-α production in inpatients, leading to a reduction of intracellular myo-inositol and its extracellular accumulation, but more in-depth analysis is required to ascertain this possibility.

In summary, our data indicate that the analysis of serum CHI3L1 and salivary myo-inositol and 2-pyrrolidineacetic acid can distinguish COVID-19 inpatients from outpatients, thus allowing avoiding unnecessary hospitalization. It is also clear that some metabolites correlate with bacteria that are found to be enriched or lost in hospitalized COVID-19 patients, suggesting that their representation may be related to bacterial taxa. Our study has some limitations. First, we excluded some metabolites because they may be related to diet differences rather than to COVID-19, even though some of these metabolites may still be important for the clinical outcome. For instance, erythritol that we found to be more represented in inpatients has been shown to enhance the virulence of some pathogens[39]; by contrast, caffeine metabolic products (theobromine, paraxanthine) that are all higher in outpatients have anti-inflammatory effects[40] and thus may be beneficial to prevent the hyperinflammatory response observed in severe COVID. Second, as COVID-19 patients were commonly treated with antibiotics, we had only 5 untreated subjects to assess the antibiotic-independent differences of microbiota composition between inpatients and outpatients. There could be other genera that characterize the two populations and that we were unable to pinpoint. Third, as our study was aimed at understanding the difference between inpatients and outpatients and the latter were kept at home, we could not follow the evolution of markers' characteristics of late responses such as interleukin 6 or IgG, and we focused on early markers' characteristics of the early response (CHI3L1 and PTX3). Fourth, the number of patients was limited, but we tried to mitigate this issue by adopting validation computational techniques and theoretical bounds that are indeed devised to deal with a limited number of observations, and we remark how the final decision three-model feature is only a two-layer deep tree, resulting in a rather low-complexity final classifier. Moreover, to support results on the separability of classes at each split, we train an SVM on the corresponding variable to estimate the "width of the margin" between the two clouds of points. Although our machine learning approach was designed to overcome the age-related bias, we cannot completely exclude that it might still affect our analysis. Nevertheless, we think that this report opens to new omic and machine learning approaches to take an informed decision on whether discharging a patient or not. This report is a proof of concept to show the power of a combined omic strategy on a biological fluid, the saliva, which is very easy to obtain, to stratify patients and to assign them to specific groups useful also for other pathologies and purposes.

Materials and Methods

Study Design

We analyzed metabolomic samples (saliva) from 320 subjects, of which 50 were COVID-19 patients (25 were treated at home—outpatients and 25 hospitalized—inpatients) and 270 were either SARS-CoV-2–naïve healthy subjects (IgG $<$12 AU/mL, n $=$ 180) or SARS-CoV-2–recovered individuals (IgG $\geq$12 AU/mL, n $=$ 90). Among SARS-CoV-2–recovered individuals,

n = 30 were asymptomatic/paucisymptomatic and n = 58 were symptomatic. Symptoms of 2 individuals were not recorded. All recovered from symptoms at the time of the sample collection. SARS-CoV-2–naïve healthy subjects (IgG <12 AU/mL, n = 180) included people with IgG ≤3.8 (n = 90) and people with IgG titers between the level of positivity (12 AU/mL) and the limit of detection of antibodies (3.8 AU/mL) whose meaning is not clear (3.8 < IgG <12 AU/mL, n = 90), and for this reason, in our analyses, we preferred to keep this group separate from the one of subjects with IgG ≤3.8. We excluded that this population represented individuals in the initial phases of an infection as all of them were negative for SARS-CoV-2 nasopharyngeal swab.[41] Demographic data of healthy controls are reported in Table A4. These observational cohort studies were conducted at Istituto Clinico Humanitas and approved by the Institutional Review Board of Istituto Clinico Humanitas. Samples were collected from Humanitas group health care employees and administrative staff (ClinicalTrials.gov NCT04451577) (individuals aged ≥18 years) and from COVID-19 patients (ClinicalTrials.gov NCT04552340) (patients who underwent to a swab or bronchoalveolar lavage test for the presence of SAR-CoV-2 infection, aged ≥18 years). All participants signed an informed consent. All the COVID-19 patients' information is shown in Table 1 and Table 2. Even though we collected samples over the course of the disease, in this study, we analyzed data from a single time point as close as possible to symptom occurrence. In particular, for COVID-19 patients and nonhospitalized employees who were SARS-Cov-2 positive (outpatients), we collected the saliva and serum as closely as possible to SARS-CoV-2 nasal swab positivity.

The primary outcome of this study was to evaluate the metabolomic profile of SARS-CoV-2 patients that required hospitalization with respect to the rest of the sample population. We analyzed also the salivary microbiota and CHI3L1 and PTX3 levels in the blood of inpatients and outpatients. No power analysis was performed to calculate the sample size. No randomization was performed.

IgG Measure

For the determination of IgG anti-SARS-CoV-2, the Liaison SARS-CoV-2 S1/S2 IgG assay (DiaSorin, Saluggia (VC), Italy) was used.[42] According to the kit manufacturer, the test discriminates among negative (<15 AU/mL; with 3.8 as the limit of IgG detection) and positive (≥15 AU/mL) subjects. However, we considered positive subjects with IgG plasma levels ≥12 AU/mL rather than those with IgG ≥15 AU/mL, as suggested by the test manufacturer, based on our previous findings showing that these two groups behaved very similarly.[41]

Metabolomic Analysis

Global metabolomic profiling was performed by Metabolon Inc. In brief, each sample received was accessioned into the Metabolon LIMS system and was assigned by the LIMS a unique identifier that was associated with the original source identifier only. Samples were prepared using an automated MicroLab STAR system (Hamilton Co). Sample preparation was conducted using methanol extractions to remove the protein fraction while allowing maximum recovery of chemically diverse metabolites. The resulting extract was divided into five fractions: two for analysis by two separate reverse phase ultra performance liquid chromatography–tandem mass spectroscopy (RP/UPLC-MS/MS) methods with positive ion mode electrospray ionization (ESI), one for analysis by RP/UPLC-MS/MS with negative ion mode ESI, one for analysis by hydrophilic interaction liquid chromatography/UPLC-MS/MS with negative ion mode ESI, and one sample was reserved for backup. All methods used Waters ACQUITY ultra performance liquid chromatography (UPLC) and a Thermo Scientific Q-Exactive high-resolution/accurate mass spectrometer interfaced with a heated electrospray ionization source and Orbitrap mass analyzer operated at 35,000 mass resolution. The sample extract was dried and then reconstituted in solvents compatible with each of the four methods. Each reconstitution solvent contained a series of standards at fixed concentrations to ensure injection and chromatographic consistency. Raw data were extracted, peak-identified, and quality control was done using Metabolon's hardware and software. Compounds were identified by comparison with library entries of purified standards or recurrent unknown entities that contain the retention time/index (RI), mass-to-charge ratio (*m/z*), and chromatographic data (including MS/MS spectral data) for all molecules present in the library. For each metabolite, the raw values in the experimental samples were divided by the median of those samples in each instrument batch, giving each batch and thus the metabolite a median of one. Batch-normalized data, simply reflecting the median-scaled raw data, were divided by the value of the normalizer. Then, each metabolite has been normalized to osmolarity and was rescaled to have median = 1 (divide the new values by the overall median for each metabolite). Data were expressed as a scaled intensity value. Welch's 2-sample t-test is used to test whether 2 unknown means are different from 2 independent populations.

Sample Collection and PTX3 Measurement

Venous blood samples were collected in BD Vacutainer Blood Collection Tubes containing ethylenediaminetetraacetic acid (EDTA), centrifuged, and stored at −80 °C until use. PTX3 plasma levels were measured by a sandwich enzyme-linked immunosorbent assay (ELISA) developed in-house, as previously described,[12] by personnel blind to patients' characteristics. For each sample, two dilutions in duplicate were tested. The assay has a detection limit of 0.1 ng/mL and an interassay variability from 8% to 10%.

CHI3L1 ELISA

Human CHI3L1 levels were measured using the DuoSet ELISA kit (DY2599, R&D systems) as per the manufacturer's instructions. Blood was always processed in less than 8 hours after collection and was kept at 4 °C to avoid nonspecific release of CHI3L1 from neutrophils.

Absorbance was measured at 450 nm using the Clariostar Plate reader (BMG Labtech).

Analysis of the Microbiota Composition by 16S rRNA Gene Sequencing

Before extraction, 200 μl of frozen saliva was thawed and heat-treated at 56 °C for 30 minutes to inactivate the live SARS-CoV-2 virus. DNA from saliva was extracted with GNOME DNA isolation kit (MPBio) following a previously published

protocol.[43] Briefly, a volume of 200 μl of inactivated saliva was added to 550 μl cell suspension solution. After addition of 50 μl RNase Mix and 100 μl cell lysis/denaturing solution, samples were incubated at 55 °C for 30 minutes. Protease Mix was added, and samples were incubated for 2 hours at 55 °C followed by a mechanical disruption step using 0.1-mm zirconia/silica beads. Lysates were retrieved, and beads were washed three times with 400 μl of TENP (50 mM **T**ris at pH 8, 20 mM **E**DTA at pH 8, 100 mM **N**aCl, 1% **P**VPP - polyvinylpolypyrrolidone) buffer. All retrieved supernatant was precipitated with isopropanol. The DNA pellet was resuspended with 400 μl water and incubated with 100 μl of salt out mixture from the GNOME DNA Kit to remove impurities. Samples were then precipitated and resuspended in water.

DNA quantity and integrity were checked through TapeStation (Agilent Technologies), and sample library preparation for next-generation sequencing was carried out using the QIAseq 16S/ITS Region Panels kit (QIAGEN), targeting the V3V4 hypervariable regions of the bacterial 16S rRNA gene and the fungal internal transcribed spacer region.

Libraries were checked through TapeStation 4200 (Agilent Technologies) and quantified by using MicroPlate Reader GloMax (Promega); libraries were then pooled at equimolar concentrations and sequenced on a MiSeq Illumina sequencer; at least 100.000 paired end reads with a length of 275 bp were produced per sample. Quality filtering of sequencing reads was executed with Trimmomatic v0.39[44] using the following parameter: AVGQUAL:30. Sequences of amplification primers and reads with more than 3 unknown (N) nucleotides were removed using cutadapt v1.18.[45] High-quality and cleaned sequences were analyzed using the Qiime2 platform (v2019.7).[46] Sequence denoise was performed separately for each sequencing batch with the qiime dada2 denoise-paired command setting the following parameters: –p-trunc-len-f 242 –p-trunc-len-r 242, and amplicon sequence variants were generated. Diversity measures (alpha- and β-diversity indices) were calculated using the qiime diversity core-metrics-phylogenetic function with a sampling depth of 23,800 sequences. Alpha diversity was evaluated by Chao1 and Shannon index and represented by the box-and-whisker plot. Community dissimilarities (β-diversity) were evaluated by Bray-Curtis distance and represented by a principal coordinates analysis plot. Differences of alpha-diversity indices across experimental groups were evaluated with the Wilcoxon rank-sum test. The qiime diversity beta-group-significance function was used to assess differences in the microbiome composition across the different experimental groups with permutational multivariate analysis of variance. Q2-feature-classifier, trained on the SILVA132 99% OTUs, specifically on the V3V4 region, was used to perform taxonomic classification. Raw taxonomic counts classified at the genus level were converted into relative abundances, and bacteria showing different abundance between the inpatient and outpatient groups were identified using the Wilcoxon rank-sum test (FDR <0.05). The magnitude of the change in abundance was expressed as a generalized fold change[47] which is calculated as the mean difference in a set of predefined quantiles of the logarithmic inpatient and outpatient distributions (quantiles ranging from 0.1 to 0.9 in increments of 0.1 were used). Spearman correlation of microbes with metabolites and serum CHI3L1 was evaluated using the corr.test function in the psych (v2.0.9) R package for each pairwise combination of microbial genus relative abundances and metabolite intensities, and the features with at least one significant correlation (FDR <0.05) were selected for downstream analysis and representation. The correlation coefficients of the 9 selected metabolites vs the microbial abundances were represented by a heatmap with hierarchical clustering generated using the pheatmap (v 1.0.12) R package. Alternatively, correlation coefficients and FDRs were represented by a bubble plot created with ggplot2 (v3.3.2) R package.

Machine Learning and Statistical Analysis

To properly extract the underlying metabolomic profile that characterizes each class of patients, a multivariate statistical analysis was performed both for feature selection and for discrimination between inpatients and outpatients.

Feature Selection. The feature selection algorithm was performed with the aim of extracting the combination of metabolites which could discriminate hospitalized patients in respect with the rest of the population.

The model has been evaluated on a training set which comprises the 80% of hospitalized patients (n = 20) and 5 patients age-matched randomly sampled from 270 SARS-CoV-2–naïve healthy subjects or SARS-CoV-2–recovered individuals and 25 COVID-19 outpatients (n = 20). The other observations were inserted in the test set.

The feature selection of the metabolite was performed by an iterative bootstrap version of the RFE algorithm, where feature ranking could be evaluated by the number of times each feature was selected in a single iteration of the RFE algorithm. The first 9 features of the ranking were selected for further steps.

Multivariate Analysis. The best 9 metabolomic features were combined together with the PCA algorithm, and the first 3 principal components were evaluated. The final supervised learning model is based on logistic regression (LR), and it was trained on the 3 principal components of the training set and validated on the test set. In addition, the probability of the prediction on the test set was computed with the LR model.

To prevent a bias due to the arbitrary decision of the training set, the previous steps were repeated several times changing the train/test set with the aim to evaluate a CI of prediction for each subject.

Finally, the best probability threshold for the LR model prediction of belonging in hospitalized or outpatient class was computed by maximizing the geometric mean (G-mean) of the receiver operating characteristic curve:

$$OptThreshold = \left\{ x \;\middle|\; max_x\left(\sqrt{TPR(x) * (1 - FPR(x))}\right) \right\}$$

Where x is the probability threshold, TPR is the true positive rate (sensitivity), and FPR is the false positive rate (specificity).

To increase the predictive power of the 9 metabolites to distinguish inpatients vs outpatients, a multivariate analysis based on the DT classifier algorithm was performed by combining metabolites and CHI3L1 and PTX3 values. DT was chosen to provide an interpretable approach to our data. As DT is a nonprobabilistic method, we also computed an SVM on each split of the DT. The strength of the boundaries was assessed by the number of support vectors required, as well as classification metrics (accuracy and area under the receiver characteristic operating curve) in a 10-fold random permutated cross-validation.

To mitigate the issue related to the small size of our data set, validation computational techniques and theoretical bounds, which are indeed devised to deal with a limited number of observations, were adopted. In particular, the final LR model is trained on the first three components of a PCA analysis ran on the 9 best metabolites. This allows us to lower the overall complexity of the final model and address possible multicollinearity problems between features. Besides, the provided metrics were obtained by adopting a cross-validation strategy which allows us to gauge the stability of our model's performance on different train-test split of the available data.

Second, the final decision 3-model feature is only a 2-layer deep tree, resulting in a rather low-complexity final classifier. Moreover, to support results on the separability of classes at each split, an SVM on the corresponding variable to estimate the "width of the margin" between the two clouds of points was trained. In more depth, as SVMs provide a theoretical bound for the difference between training and test performance which depends on the number of the support points, class separation at each decision split by evaluating the number of such points was assessed. Given the low number of support vectors (Figure 2E) with respect to the number of subjects of our cohort, we are confident on the stability of our results.

All authors had access to the study data and had reviewed and approved the final manuscript.

References

1. Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 2020;181(2):271–280.
2. Ou X, Liu Y, Lei X, et al. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nat Commun 2020;11(1):1620.
3. Arunachalam PS, Scott MKD, Hagan T, et al. Systems vaccinology of the BNT162b2 mRNA vaccine in humans. Nature 2021;596(7872):410–416.
4. Blasco H, Bessy C, Plantier L, et al. The specific metabolome profiling of patients infected by SARS-COV-2 supports the key role of tryptophan-nicotinamide pathway and cytosine metabolism. Sci Rep 2020;10(1):16824.
5. Bruzzone C, Bizkarguenaga M, Gil-Redondo R, et al. SARS-CoV-2 infection dysregulates the metabolomic and lipidomic profiles of serum. iScience 2020;23(10):101645.
6. Barberis E, Timo S, Amede 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 2020;21(22):8623.
7. Shen B, Yi X, Sun Y, et al. Proteomic and metabolomic characterization of COVID-19 patient sera. Cell 2020;182(1):59–72 e15.
8. Song JW, Lam SM, Fan X, et al. Omics-driven systems interrogation of metabolic dysregulation in COVID-19 pathogenesis. Cell Metab 2020;32(2):188–202 e5.
9. Iebba V, Zanotta N, Campisciano G, et al. Profiling of oral microbiota and cytokines in COVID-19 patients. Front Microbiol. 2021;12:671813.
10. Kamle S, Ma B, He CH, et al. Chitinase 3-like-1 is a therapeutic target that mediates the effects of aging in COVID-19. JCI Insight 2021;6(21):e148749.
11. Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. Machine Learn 2002;46(1):389–422.
12. Brunetta E, Folci M, Bottazzi B, et al. Macrophage expression and prognostic significance of the long pentraxin PTX3 in COVID-19. Nat Immunol 2021;22(1):19–24.
13. Lian M, Cao H, Baranova A, et al. Aging-associated genes TNFRSF12A and CHI3L1 contribute to thyroid cancer: an evidence for the involvement of hypoxia as a driver. Oncol Lett 2020;19(6):3634–3642.
14. Reynolds LA, Finlay BB. A case for antibiotic perturbation of the microbiota leading to allergy development. Expert Rev Clin Immunol 2013;9(11):1019–1030.
15. Ubeda C, Pamer EG. Antibiotics, microbiota, and immune defense. Trends Immunol 2012;33(9):459–466.
16. He X, McLean JS, Edlund A, et al. Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. Proc Natl Acad Sci USA 2015;112(1):244–249.
17. To KK, Tsang OT, Leung WS, et al. Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study. Lancet Infect Dis 2020;20(5):565–574.
18. Ali F, Sweeney DA. No one likes a stick up their nose: making the case for saliva-based testing for coronavirus disease 2019 (COVID-19). Clin Infect Dis 2021;72(9):e357–e358.
19. Kwon T, Gaudreault NN, Richt JA. Seasonal stability of SARS-CoV-2 in biological fluids. Pathogens 2021;10(5).
20. Chen Y, Gu S, Chen Y, et al. Six-month follow-up of gut microbiota richness in patients with COVID-19. Gut 2021;71(1):222–225.
21. Lv L, Jiang H, Chen Y, et al. The faecal metabolome in COVID-19 patients is altered and associated with clinical features and gut microbes. Anal Chim Acta 2021;1152:338267.
22. Zuo T, Zhang F, Lui GCY, et al. Alterations in gut microbiota of patients with COVID-19 during time of hospitalization. Gastroenterology 2020;159(3):944–955.e8.
23. Lv L, Gu S, Jiang H, et al. Gut mycobiota alterations in patients with COVID-19 and H1N1 infections and their associations with clinical features. Commun Biol 2021;4(1):480.
24. Ventero MP, Cuadrat RRC, Vidal I, et al. Nasopharyngeal microbial communities of patients infected with SARS-CoV-2 that developed COVID-19. Front Microbiol 2021;12:637430.
25. Rosas-Salazar C, Kimura KS, Shilts MH, et al. SARS-CoV-2 infection and viral load are associated with the upper respiratory tract microbiome. J Allergy Clin Immunol 2021;147(4):1226–1233.e2.
26. Sulaiman I, Chung M, Angel L, et al. Microbial signatures in the lower airways of mechanically ventilated COVID-19

patients associated with poor clinical outcome. Nat Microbiol 2021;6(10):1245–1258.
27. Xu R, Lu R, Zhang T, et al. Temporal association between human upper respiratory and gut bacterial microbiomes during the course of COVID-19 in adults. Commun Biol 2021;4(1):240.
28. Martino C, Kellman BP, Sandoval DR, et al. Bacterial modification of the host glycosaminoglycan heparan sulfate modulates SARS-CoV-2 infectivity. bioRxiv 2020. http://doi.org/10.1101/2020.08.17.238444.
29. Firpo MR, Mastrodomenico V, Hawkins GM, et al. Targeting polyamines inhibits coronavirus infection by reducing cellular attachment and entry. ACS Infect Dis 2020;7(6):1423–1432.
30. Tuomi K, Raina A, Mantyjarvi R. 1,3-Diaminopropane rapidly inhibits protein synthesis and virus production in BKT-1 cells. FEBS Lett 1980;111(2):329–332.
31. Brinig MM, Lepp PW, Ouverney CC, et al. Prevalence of bacteria of division TM7 in human subgingival plaque and their association with disease. Appl Environ Microbiol 2003;69(3):1687–1694.
32. Podar M, Abulencia CB, Walcher M, et al. Targeted access to the genomes of low-abundance organisms in complex microbial communities. Appl Environ Microbiol 2007;73(10):3205–3214.
33. Bor B, Poweleit N, Bois JS, et al. Phenotypic and physiological characterization of the epibiotic interaction between TM7x and its basibiont actinomyces. Microb Ecol 2016;71(1):243–255.
34. Zhang G, Tian Y, Hu K, et al. Importance and regulation of inositol biosynthesis during growth and differentiation of Streptomyces. Mol Microbiol 2012;83(6):1178–1194.
35. Roberts MF. Inositol in bacteria and archaea. Subcell Biochem 2006;39:103–133.
36. Hallman M, Järvenpää AL, Pohjavuori M, et al. Myoinositol supplementation reduces severe complications of the respiratory distress syndrome (RDS). Pediatr Res 1985;19(10):1077.
37. Hallman M, Spragg R, Harrell JH, et al. Evidence of lung surfactant abnormality in respiratory failure. Study of bronchoalveolar lavage phospholipids, surface activity, phospholipase activity, and plasma myoinositol. J Clin Invest 1982;70(3):673–683.
38. Yorek MA, Dunlap JA, Thomas MJ, et al. Effect of TNF-alpha on SMIT mRNA levels and myo-inositol accumulation in cultured endothelial cells. Am J Physiol 1998;274(1):C58–C71.
39. Petersen E, Rajashekara G, Sanakkayala N, et al. Erythritol triggers expression of virulence traits in Brucella melitensis. Microb Infect 2013;15(6-7):440–449.
40. Barcelos RP, Lima FD, Carvalho NR, et al. Caffeine effects on systemic metabolism, oxidative-inflammatory pathways, and exercise performance. Nutr Res 2020;80:1–17.
41. Sandri MT, Azzolini E, Torri V, et al. SARS-CoV-2 serology in 4000 health care and administrative staff across seven sites in Lombardy, Italy. Sci Rep. 2021;11(1):12312.
42. Bonelli F, Sarasini A, Zierold C, et al. Clinical and analytical performance of an automated serological test that identifies S1/S2-neutralizing IgG in COVID-19 patients semiquantitatively. J Clin Microbiol 2020;58(9):e01224–20.
43. Furet J-P, Firmesse O, Gourmelon M, et al. Comparative assessment of human and farm animal faecal microbiota using real-time quantitative PCR. FEMS Microbiol Ecol 2009;68(3):351–362.
44. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15):2114–2120.
45. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads 2011;7(1):3.
46. Bolyen E, Rideout JR, Dillon MR, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 2019;37(8):852–857.
47. Wirbel J, Pyl PT, Kartal E, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med 2019;25(4):679–689.

Received September 24, 2021. Accepted December 6, 2021.

Correspondence:
Address correspondence to: Prof. Maria Rescigno, PhD, Humanitas University Pieve Emanuele, Milan, Italy. e-mail: maria.rescigno@hunimed.eu.

Acknowledgments:
The authors would like to thank all the employees and the patients who volunteered to participate to this study and all the nurses and personnel who collected the samples.

Author's Contributions:
Conceptualization: Maria Rescigno, Chiara Pozzi, Giuseppe Penna, Riccardo Levi, Daniele Braga; Methodology: Daniele Braga, Riccardo Levi, Francesco Carli, Abbass Darwich, Ilaria Spadoni, Bianca Oresta, Carola Conca Dioguardi, Leonardo Ubaldi, Giovanni Angelotti, ICH COVID-19 Task-force; Investigation: Chiara Pozzi, Riccardo Levi, Daniele Braga, AbD, AnD, Giuseppe Penna, Maria Rescigno; Visualization: Chiara Pozzi, Riccardo Levi, Daniele Braga, Maria Rescigno; Funding acquisition: Maria Rescigno, Alberto Mantovani; Project administration: Elena Azzolini, Maurizio Cecconi; Supervision: Barbara Bottazzi, Cecilia Garlanda, Antonio Voza, Alberto Mantovani, Riccardo Barbieri, Letterio S. Politi; Writing – original draft: Maria Rescigno; Writing – review & editing: Maria Rescigno, Chiara Pozzi, Maurizio Cecconi, Daniele Braga, Riccardo Levi. All authors had access to the study data and had reviewed and approved the final manuscript.

Conflicts of Interest:
The authors disclose no conflicts.

Funding:
The study was funded by Italian Ministry of Health Ricerca Finalizzata COVID-2020-12371640 (AM, MR) and Italian Ministry of Health Ricerca Corrente (AM, MR).

Ethical Statement:
The corresponding author, on behalf of all authors, jointly and severally, certifies that their institution has approved the protocol for any investigation involving humans and that all experimentation was conducted in conformity with ethical and humane principles of research.

Data Transparency Statement:
The deidentified individual participant data and the code are deposited in Institutional Zenodo community named IRCCS Humanitas Research Hospital & Humanitas University. They are available at the link https://zenodo.org/record/5151211#.YQZcW0DONaQ with restricted license, however available upon request. Patient informed consent does not allow for the deposition of clinical data in public access repositories. Interested researchers should contact biblioteca@humanitas.it to inquire about access; requests for noncommercial academic use will be considered and require ethics review. 16S rRNA data are available online at the Sequence Read Archive under the BioProject accession number PRJNA750902 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA750902).