nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-26500-8 OPEN

Clinical practices underlie COVID-19 patient respiratory microbiome composition and its interactions with the host

Verónica Lloréns-Rico [1,2], Ann C. Gregory[1,2], Johan Van Weyenbergh[3], Sander Jansen [4], Tina Van Buyten[4], Junbin Qian [5,6], Marcos Braz[3], Soraya Maria Menezes[3], Pierre Van Mol [7,8,9], Lore Vanderbeke [10], Christophe Dooms[9,11], Jan Gunst [12], Greet Hermans[12], Philippe Meersseman[13], CONTAGIOUS collaborators*, Els Wauters[9,11], Johan Neyts [4], Diether Lambrechts [7,8], Joost Wauters[13,17] & Jeroen Raes [1,2,17]✉

Understanding the pathology of COVID-19 is a global research priority. Early evidence suggests that the respiratory microbiome may be playing a role in disease progression, yet current studies report contradictory results. Here, we examine potential confounders in COVID-19 respiratory microbiome studies by analyzing the upper ($n = 58$) and lower ($n = 35$) respiratory tract microbiome in well-phenotyped COVID-19 patients and controls combining microbiome sequencing, viral load determination, and immunoprofiling. We find that time in the intensive care unit and type of oxygen support, as well as associated treatments such as antibiotic usage, explain the most variation within the upper respiratory tract microbiome, while SARS-CoV-2 viral load has a reduced impact. Specifically, mechanical ventilation is linked to altered community structure and significant shifts in oral taxa previously associated with COVID-19. Single-cell transcriptomics of the lower respiratory tract of COVID-19 patients identifies specific oral bacteria in physical association with proinflammatory immune cells, which show higher levels of inflammatory markers. Overall, our findings suggest confounders are driving contradictory results in current COVID-19 microbiome studies and careful attention needs to be paid to ICU stay and type of oxygen support, as bacteria favored in these conditions may contribute to the inflammatory phenotypes observed in severe COVID-19 patients.

[1] Laboratory of Molecular Bacteriology, Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium. [2] Center for Microbiology, VIB, Leuven, Belgium. [3] Laboratory for Clinical and Evolutionary Virology, Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium. [4] Laboratory of Virology and Chemotherapy, Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium. [5] Zhejiang Provincial Key Laboratory of Precision Diagnosis and Therapy for Major Gynecological Diseases, Women's Hospital, Zhejiang University School of Medicine, Hangzhou, China. [6] Institute of Genetics, Zhejiang University School of Medicine, Hangzhou, China. [7] Laboratory of Translational Genetics, Department of Human Genetics, KU Leuven, Leuven, Belgium. [8] VIB Center for Cancer Biology, VIB, Leuven, Belgium. [9] Department of Pneumology, University Hospitals Leuven, Leuven, Belgium. [10] Laboratory of Clinical Bacteriology and Mycology, Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium. [11] Laboratory of Respiratory Diseases and Thoracic Surgery (BREATHE), Department of Chronic Diseases and Metabolism, KU Leuven, Leuven, Belgium. [12] Laboratory of Intensive Care Medicine, Department of Cellular and Molecular Medicine, KU Leuven, Leuven, Belgium. [13] Laboratory for Clinical Infectious and Inflammatory Disorders, Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium. [17]These authors contributed equally: Joost Wauters, Jeroen Raes. *A list of authors and their affiliations appears at the end of the paper. ✉email: jeroen.raes@kuleuven.vib.be

COVID-19, a novel coronavirus disease classified as a pandemic by the World Health Organization, has caused over 150 million reported cases and 3 million deaths worldwide to date. Infection by its causative agent, the novel coronavirus SARS-CoV-2, results in a wide range of clinical manifestations: it is estimated that around 80% of infected individuals are asymptomatic or present only mild respiratory and/or gastrointestinal symptoms, while the remaining 20% develop acute respiratory distress syndrome requiring hospitalization and oxygen support and, of those, 25% of cases necessitate critical care. Despite a concerted global research effort, many questions remain about the full spectrum of the disease severity. Independent studies from different countries, however, agree that age and sex are the major risk factors for disease severity and patient death[1–3], as well as type 2 diabetes and obesity[4,5]. Other potential risk factors for critical condition and death are a viral load of the patient upon hospital admission[6–8] and the specific immune response to infection, with the manifestation of an abnormal immune response in critical patients characterized by dysregulated levels of proinflammatory cytokines and chemokines, which some studies have associated with organ failure[9,10].

Despite its close interplay with the immune system and its known associations with host health, little is known about the role of the respiratory microbiota in modulating COVID-19 disease severity, or its potential as a prognostic marker[11]. Previous studies exploring other pulmonary disorders have shown that the lung microbiota members may exacerbate symptoms and contribute to their severity[12], potentially through direct crosstalk with the immune system and/or due to bacteremia and secondary infections[13]. First studies of the respiratory microbiome in COVID-19 have revealed elevated levels of opportunistic pathogenic bacteria[14–16]. However, reports on bacterial diversity are contradictory. While some studies report a low microbial diversity in COVID-19 patients[14,17] that rebounds following recovery[15], others show an increased diversity in the COVID-19 associated microbiota[16]. These conflicting results could be due to differences in sampling location (upper or lower respiratory tract), the severity of the patients, disease stage, treatment, or other confounders. While these early findings already suggest that the lung microbiome could be exacerbating or mitigating COVID-19 progression, exact mechanisms are yet to be elucidated. Therefore, an urgent need exists for studies identifying and tackling confounders in order to discern true signals from noise.

To identify potential associations between the COVID-19 severity and evolution and the upper and lower respiratory tract microbiota, we used nasopharyngeal swabs and bronchoalveolar lavage (BAL) samples, respectively. For the upper respiratory tract, we longitudinally profiled the nasopharyngeal microbiome of 58 COVID-19 patients during intensive care unit (ICU) treatment and after discharge to a classical hospital ward following clinical improvement, in conjunction with viral load determination and nCounter immune profiling. For the lower respiratory tract, we profiled microbial reads in cross-sectional single-cell RNA-seq data[18] from bronchoalveolar lavage (BAL) samples of 22 COVID-19 patients and 13 pneumonitis controls with negative COVID-19 qRT-PCR, obtained from the same hospital. The integration of these data enabled us to (1) identify potential confounders of COVID-19 microbiome associations, (2) explore how microbial diversity evolves throughout hospitalization, (3) study microbe-host cell interactions, and (4) substantiate a link between the respiratory microbiome and SARS-CoV-2 viral load, as well as COVID-19 disease severity. Altogether, our results suggest the existence of associations between the microbiota and specific immune cells in the context of COVID-19 disease. These interactions may be driven by mechanical ventilation and its associated clinical practices, and therefore could potentially influence COVID-19 disease progression and resolution.

Results

The upper respiratory microbiota of COVID-19 patients. We longitudinally profiled the upper respiratory microbiota of 58 patients diagnosed with COVID-19 based on a positive qRT-PCR test or a negative test with high clinical suspicion based on symptomatology and a chest CT scan showing typical round glass opacities. All these patients were admitted and treated at UZ Leuven hospital. Patient demographics for this cohort are shown in Table 1.

In total, 112 nasopharyngeal swabs from these patients were processed (Fig. 1a): the V4 region of the 16S rRNA gene was amplified on extracted DNA using 515F and 806R primers and sequenced on an Illumina MiSeq platform (see Methods). From the same swabs, RNA was extracted to determine SARS-CoV-2 viral loads and to estimate immune cell populations of the host and expression of immune-related genes using nCounter (Methods). Of the 112 samples processed and sequenced, 101 yielded over 10,000 amplicon reads that could be assigned to bacteria at the genus level (Fig. 1b; Methods). The microbiome of the entire cohort was dominated by the gram-positive genera *Staphylococcus* and *Corynebacterium*, typical from the nasal cavity and nasopharynx[19].

Bacterial alpha diversity is associated with ICU stay length, SARS-CoV-2 viral load, and calprotectin levels. First, we determined genus-level alpha diversity for the 101 samples with more than 10,000 genus-level assigned reads, using the Shannon Diversity index (SDI; see Methods; Supplementary Table 1). We observed that the SDI was significantly different across sampling moments (Kruskal–Wallis test, p value = 0.009; Supplementary Figure 1a), with significant differences between the swabs procured upon patient ICU admission and later timepoints, suggesting an effect of disease progression and/or treatment (for instance due to antibiotics administered throughout ICU stay). We explored these differences further, and observed that SDI correlated negatively with the number of days spent in ICU at the

Table 1 Patient demographics of upper and lower respiratory tract cohorts.

	Upper respiratory tract (swabs)	Lower respiratory tract (BAL)
Number of patients	58	35
Type of sampling	Longitudinal	Cross-sectional
COVID-19 diagnosis (%)	58 (100%)	22 (63%)
Patients admitted to ICU (%)	58 (100%)	21 (60%) at sampling
Age (range)	61 (37-83)	64 (45-85)
Female sex (%)	13 (22%)	12 (34%)
BMI (range)	29 (22-47)	26 (16-36)
Diabetic (%)	12 (21%)	6 (17%)
Days in ICU (range)	21.4 (2-72)	Not Available (cross-sectional cohort)
Days in hospital (range)	32.5 (6-86)	Not Available (cross-sectional cohort)

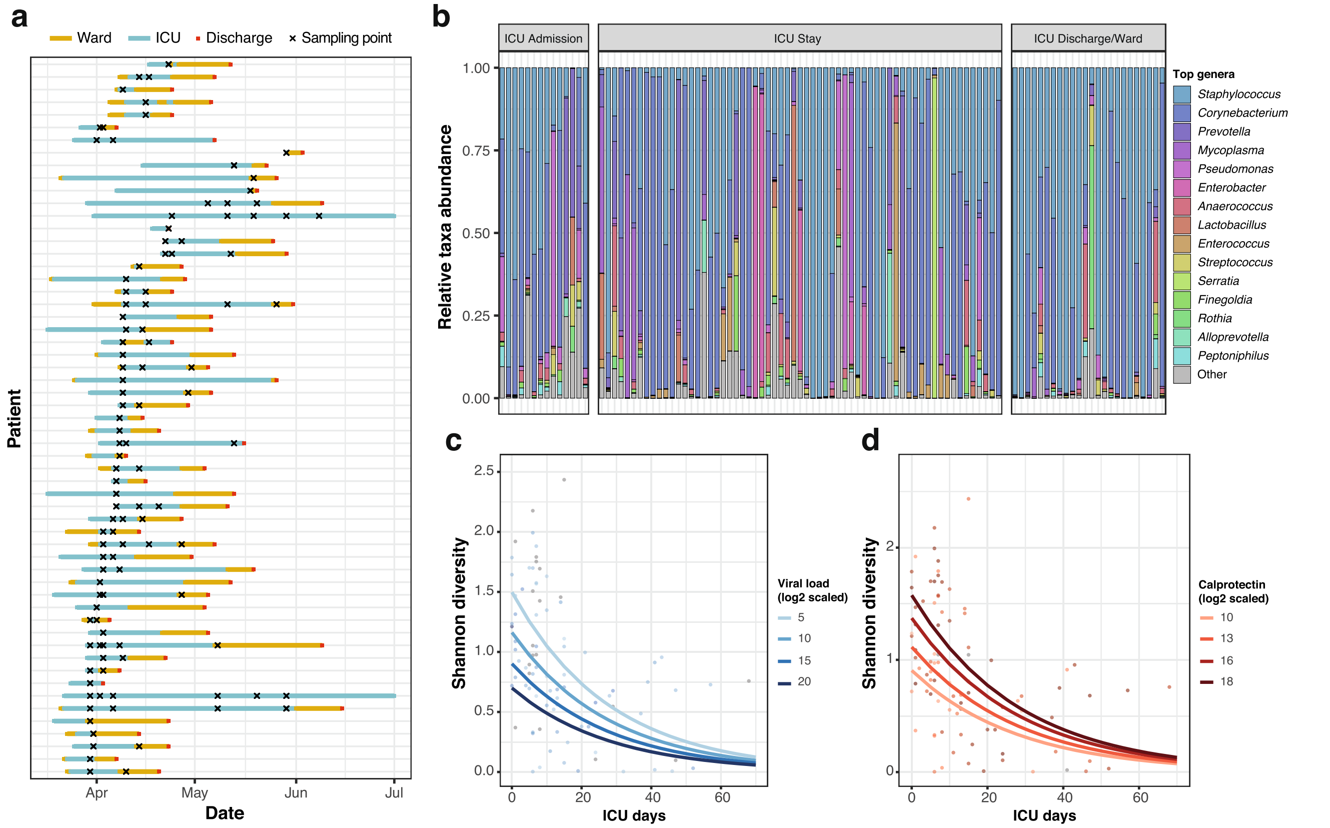

Fig. 1 Sample overview and alpha diversity. a Longitudinal sampling of patients ($n = 58$ patients). Each line represents one patient. Yellow lines span the days spent in ward, while blue lines span the days spent in ICU. Red points mark hospital discharge dates. Crosses indicate the timepoints where swab samples were obtained for microbiome analyses. **b** Top 15 most abundant genera in this cohort. Samples with >10,000 reads assigned to microbial taxa at the genus level ($n = 101$) were stratified according to the sampling moment: upon admission, throughout the ICU stay or at ICU discharge/during treatment in ward. **c** Effect of the length of ICU stay and SARS-CoV-2 viral load on upper respiratory tract microbiome diversity ($n = 89$, after removal of samples with missing data). The plot shows the model-predicted Shannon index as a function of the days in ICU, for different levels of SARS-CoV-2 viral load (selected within the range of observed data). Confidence intervals for the predictions are shown in Supplementary Figure 1c. **d** Association of the length of ICU stay and calprotectin gene expression levels with upper respiratory tract microbiome diversity ($n = 89$, after removal of samples with missing data). The plot shows the model-predicted Shannon index as a function of the days in ICU, for different levels of calprotectin RNA (subunit S100A8) gene expression, selected within the range of observed data. Confidence intervals for the predictions are shown in Supplementary Figure 1d. Source data are provided as a Source Data file.

moment of sampling, with longer ICU stays leading to a lower diversity ($\rho = -0.53$, p value $= 1.9 \times 10^{-8}$).

To evaluate the association of other clinical or disease-related variables with upper respiratory tract microbiome diversity, we used a generalized linear mixed model framework: we performed an exhaustive screening of all possible models containing up to eight different explanatory variables, using an automated model selection algorithm (see Methods). The variables used to regress the SDI comprise the patient ID, modeled as a random effect; disease-related variables, such as the time in ICU, SARS-CoV-2 viral load or the use of mechanical ventilation; and other variables known to affect the microbiome, such as the administration of antibiotics (specifically meropenem/piperacillin-tazobactam and ceftriaxone) or the levels of inflammatory markers (calprotectin, C-reactive protein). The antibiotics meropenem and piperacillin/tazobactam were grouped as a single variable in all subsequent analyses as they were administered under the same clinical guidelines. The best performing model (AICc = 121.79; p value $= 4.06 \times 10^{-8}$) included the patient modeled as a random effect and confirmed a negative association between the time spent in ICU and diversity. Additionally, this model showed a negative effect of SARS-CoV-2 viral load and a positive association of calprotectin levels with the SDI (Fig. 1c, d; Supplementary Figure 1b–d).

We leveraged all the models generated in the screening to calculate weighted importance scores for all the fixed effects tested (Methods; Supplementary Figure 1e). These scores showed that the three variables incorporated in the best model (time in ICU, SARS-CoV-2 viral load, and calprotectin) held the highest relative importance, followed by CRP levels and mechanical ventilation. Treatment with antibiotics ceftriaxone an meropenem or piperacillin-tazobactam had the lowest importance scores, and no significant differences in SDI were found between the samples obtained before and after the administration of meropenem/piperacillin-tazobactam (Supplementary Figure 1f).

Altogether, our data suggest that respiratory microbiome diversity is linked to the length of ICU stay, SARS-CoV-2 viral load, and calprotectin levels. While no significant effects were found for the most widely used antibiotics in this cohort, we cannot rule out that antibiotic administration or other clinical practices are causing the decrease of SDI over time.

Respiratory microbiome composition variation is linked to respiratory support and associated clinical practices. We next explored potential associations between the upper respiratory genus-level microbiota composition and the extensive metadata

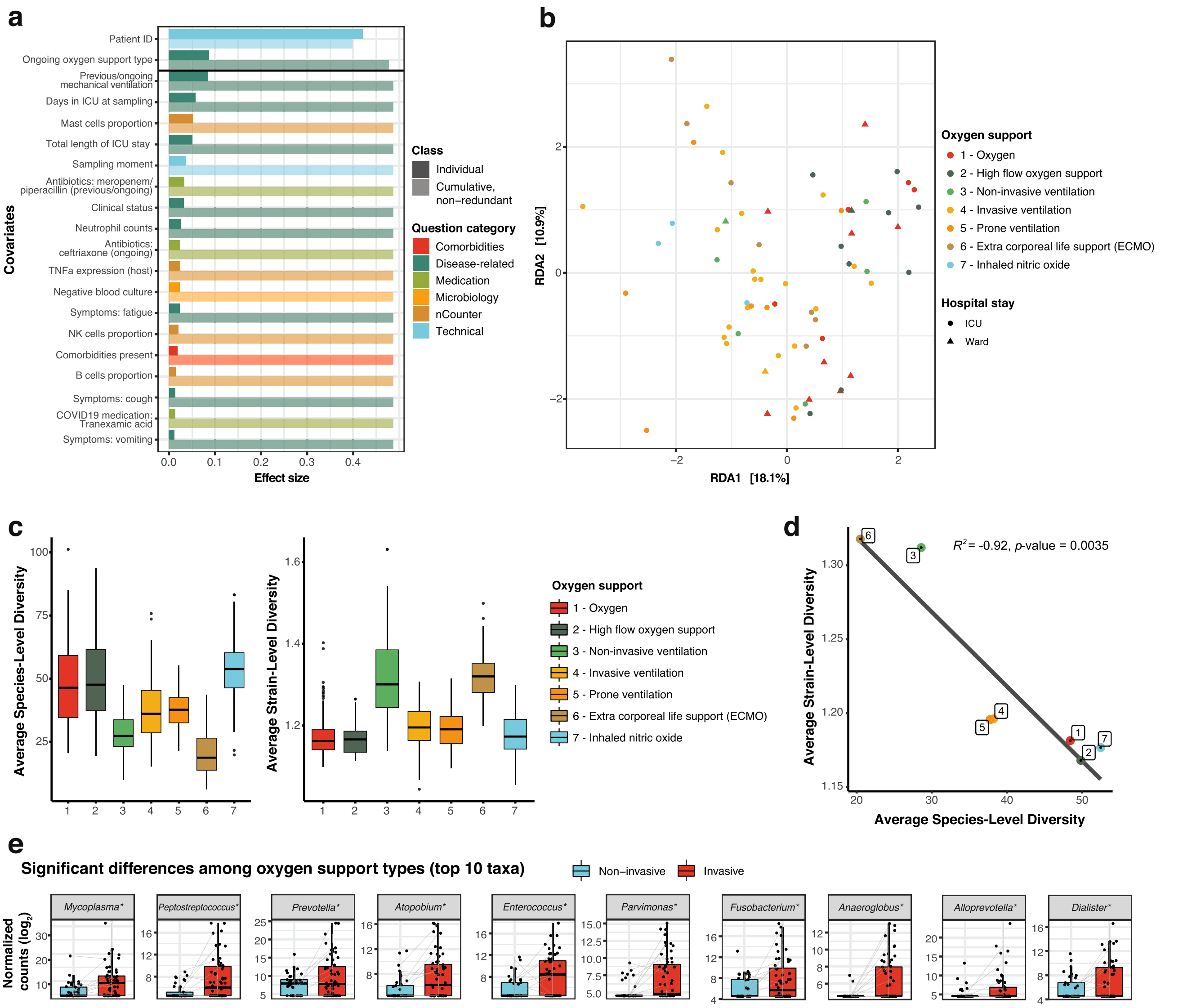

Fig. 2 Upper respiratory microbiome covariates in COVID-19. a Significant (BH-corrected *p* value < 0.05) covariates explaining microbiota variation in the upper respiratory tract in this cohort ($n = 101$ samples, dbRDA). Individual covariates are listed on the y-axis, their color corresponds to the metadata category they belong to. Darker colors refer to the individual variance explained by each of these covariates assuming independence, while lighter colors represent the cumulative and nonredundant variance explained by incorporating each variable to a model using a stepwise dbRDA analysis. The black horizontal line separates those variables that are significant in the nonredundant analysis on top (Patient ID and oxygen support type) from the rest. **b** RDA ordination plot showing the first two constrained axes. Ordination is constrained by the two significant variables "Patient ID" and "Oxygen support". Samples are depicted as points, whose color indicates the oxygen support type of the patient and whose shape indicates stay at ward or ICU (during sampling). Axes indicate the variance explained by the first two constrained components of the dbRDA analysis. **c** Species- (left) and strain-level diversity (right) of the samples, stratified by oxygen support type ($n = 101$ samples). **d** Pearson correlation between average species- and strain- level diversity for each of the oxygen support categories. The *R*2 and the *p* value (two-sided *F*-test) from a linear regression between the two variables are shown. **e**. Significant differences in taxa abundances among oxygen support types ($n = 101$ samples). Differentially abundant taxa between invasive (red) and noninvasive (blue) ventilated samples. Only the top 10 most significant taxa are shown, as determined by their BH-adjusted *p* value. Boxplots span from the first until the third quartile of the data distribution, and the horizontal line indicates the median value of the data. The whiskers extend from the quartiles until the last data point within 1.5 times the interquartile range, with outliers beyond. Individual data points are also represented. Gray lines join samples pertaining to the same patient, taken at different timepoints. Asterisks (*) indicate taxa that remain significant after controlling for the main antibiotics (ceftriaxone and meropenem/piperacillin-tazobactam). Source data are provided as a Source Data file.

collected in the study. In total, 72 covariates related to patient anthropometrics, medication, and clinical variables measured in the hospital, as well as SARS-CoV-2 viral load, host cytokine expression, and estimated immune cell populations measured in the swabs were tested (Supplementary Data 1). Individually, 20 of these covariates showed a significant correlation to microbiota composition in a univariate analysis (dbRDA, p value < 0.05; FDR < 0.05; Fig. 2a). These significant covariates were related to disease and measures of its severity, such as the clinical evaluation of the patient, the total length of the ICU stay, the number of days in ICU at the time of sampling, or the type of oxygen support needed by the patient. Despite showing an association to the overall diversity, SARS-CoV-2 viral load detected in the swabs was not significantly associated to microbiome composition variation (Supplementary Data 1). Neither ongoing antibiotic usage (i.e., administration of any type of antibiotic) nor number of ongoing antibiotics administered were significant, but the administration of specific antibiotics meropenem/piperacillin-tazobactam (previous or ongoing treatment) and ceftriaxone (ongoing administration only) showed significant associations with microbiome composition (Supplementary Data 1, Fig. 2a).

Of the 20 significant covariates, only two accounted for 48.7% nonredundant variation in this dataset in a multivariate analysis (dbRDA; p value = 0.001), with the rest holding redundant information. These were the patient ID, included due to the longitudinal sampling of patients, and confirming that intra-individual variation over time is smaller than patient inter-individual variation[20], and the type of oxygen support received at the time of sampling (Fig. 2a, b). Notably, the type of oxygen support discriminated samples based on ventilation type, with noninvasive ventilation samples (groups 1, 2, and 3) separating from samples from intubated patients (groups 4 to 7; PERMANOVA test, $R^2 = 0.0642$, p value = 0.001). Because of this separation, we also evaluated whether previous mechanical ventilation (regardless of the specific group) had a significant impact on the microbiome composition, showing even a larger effect size than when considering only the ongoing mechanical ventilation (PERMANOVA test, $R^2 = 0.0965$, p value = 0.001), suggesting that this invasive procedure may have an effect that is prolonged in time.

Mechanical ventilation is inherently associated to additional clinical practices, such as administration of broad-spectrum antibiotics and decontamination procedures (including chlorhexidine washes) to prevent/treat ventilator-associated pneumonia. Hence, we explored whether antibiotic usage could explain the significant relationship between microbiome composition and oxygen support type. We found that from the specific antibiotics associated to microbiome composition, ceftriaxone was predominantly administered in patients on noninvasive oxygen support (Chi-square, p value = 0.001), whilst meropenem or piperacillin-tazobactam were preferentially given to patents on mechanical ventilation (Chi-square test, p value = 0.002; Supplementary Figure 2a). This association is not casual and responds to current treatment guidelines at UZ Leuven University Hospital: ceftriaxone is administered to patients upon admission and for 3–7 days to prevent potential bacterial co-infections. In our cohort, 80% of the patients received ceftriaxone at the beginning of their stay (Supplementary Figure 2b). Patients with longer ICU stays and requiring higher levels of oxygen support will be considered to have hospital-acquired/ventilator-associated pneumonia (HAP/VAP) and receive meropenem or a combination of piperacillin/tazobactam (Supplementary Figure 2b). Therefore, the observed correlation between oxygen support types and these antibiotics can be explained by disease severity and length of ICU stay.

We therefore explored whether we could observe an effect of oxygen support type alone, deconfounding for the patient ID and the two significant antibiotic covariates using partial dbRDA to extract the effect size of oxygen support alone. The deconfounded model exhibited a significant association to overall microbiome composition (partial dbRDA; $R^2 = 0.058$, p value = 0.042) suggesting that although antibiotic administration may explain part of the variation in microbiome composition observed, there may an independent effect of the oxygen support type. Nevertheless, the effect of other practices concomitant to mechanical ventilation, such as oral decontamination with chlorhexidine washes, could not be disentangled as these treatments were always performed together.

To determine if oxygen support or associated practices also impacted the microbiome at finer taxonomic resolution, we revisited alpha diversity at species- and strain-level. We defined species as 97% identity 16S OTUs and strains per species as the clustered 16S sequences within each OTU. Our analyses revealed both species- and strain-level diversity change with ventilation, even with noninvasive ventilation (e.g., BIPAP, CPAP). Across all samples we observed high species- and low strain-level diversity preventilation, which reversed following any form of ventilation (Fig. 2c; Wilcoxon test; p values < 0.05, with the exception of type 7), with the exception of ventilation with inhaled nitric oxide. Further, species- and strain-level diversity showed a strong inverse correlation (Fig. 2d; Pearson's correlation, $R^2 = -0.92$, p value = 0.0035).

Given the observed effect of mechanical ventilation on the overall microbiome composition, we evaluated which specific taxa were differentially abundant between samples from intubated and nonintubated patients. In total, 28 genera were more abundant in samples from mechanically ventilated patients, while 1 genus was more abundant in non-invasively ventilated patients (p value < 0.05; FDR < 0.05; Fig. 2e, Supplementary Figure 3a; Supplementary Data 2). When controlling for the effect of the antibiotics ceftriaxone and meropenem/piperacillin, 20 genera were significantly different between both groups of samples (Supplementary Figure 3b, Supplementary Data 2). Some of these taxa are common oral microbiome commensals or opportunistic pathogens that had been repeatedly reported as more abundant in COVID-19 patients than in controls, such as *Prevotella*, *Fusobacterium*, *Porphyromonas*, or *Lactobacillus*[14–16]. Here, we reported higher abundance of these genera in mechanically ventilated COVID-19 patients as compared to nonmechanically ventilated COVID-19 patients. This points at mechanical ventilation (and associated practices such as oral decontamination) as a potential confounder of previous COVID-19 studies. Additionally, we found other taxa not previously reported in previous COVID-19 microbiome studies, such as *Mycoplasma* or *Megasphaera* (Fig. 2e, Supplementary Figure 2), but previously associated to the risk of ventilator-associated pneumonia[21].

By extracting the amplicon sequence variants (ASVs) corresponding to these differentially abundant genera (see Methods), some of these taxa could be narrowed down to the species level, confirming their origin as typically oral bacteria: for instance, *Prevotella* species included *P. oris*, *P. salivae*, *P. denticola*, *P. buccalis*, and *P. oralis*. Within the *Mycoplasma* genus, ASVs were assigned to *Mycoplasma salivarium* among other species, an oral bacterium which has been previously associated to the incidence of ventilator-associated pneumonia[21]. When controlling for ventilation type, no taxa were found associated to SARS-CoV-2 viral loads (Supplementary Data 2). These results show that further research with larger cohorts and controlling for the relevant confounders highlighted here, such as ventilation type, antibiotic usage or length of stay in ICU, will be needed to study the specific effect of the viral infection.

Single-cell RNA-seq of bronchoalveolar fluid identifies oral commensals and opportunist pathogens in the lower respiratory tract. Next, we explored what the functional consequences of (disease and/or treatment-driven) lung microbiome disturbances could be. To do so, we screened host single-cell RNA-seq data generated on BAL samples of 35 patients[18] using a computational pipeline to identify microbial reads (see Methods). All patients in this cross-sectional cohort showed clinical symptoms of pneumonia, 22 of them being diagnosed with COVID-19. The other 13 patients with non-COVID-19 pneumonia were hereafter referred to as controls (Table 1). Out of the 35 patients, 21 were admitted to ICU (20 COVID-19 patients and 1 control) and 14 were hospitalized in ward at the moment of sampling (2 COVID-19 patients and 12 controls; Table 1 and Supplementary Figure 4). Microbial read screening of these samples revealed an average of 7295.3 microbial reads per sample (ranging from 0 to 74,226 reads, with only a single sample yielding zero microbial reads; Supplementary Fig. 4).

Among the top taxa encountered in these patients, we found similarities with the data obtained in nasopharyngeal swabs. The top 15 species detected include *Mycoplasma salivarium* as the dominating taxon in five COVID-19 patients in ICU, as well as different *Prevotella* members. Non-COVID-19 pneumonia patients in ward (i.e., nonmechanically ventilated) harbored different microbes: two patients had a microbiome dominated by *Porphyromonas gingivalis*, while a single patient had a microbiome dominated by the fungus *Pneumocystis jirovecii*, a known pathogen causing *Pneumocystis* pneumonia (PCP)[22].

Bacteria in the lower respiratory tract associate to host cells from the innate immune system in COVID-19 patients. Next, we took advantage of the single-cell barcoding and questioned whether the microbial reads that we identified were found in association with host cells (for instance infecting or internalized), or contrarily, had unique barcodes suggesting a free-living state. In total, 29,886 unique barcodes were identified that matched a total of 46,151 microbial UMIs. The distribution of UMIs per barcode was asymmetrical, ranging from 1 to 201 and with 88% of the barcodes having a single UMI. Additionally, 26,572 barcodes (89%) were associated to a single microbial species, the rest being associated to 2 species (8.8%) or more (2.2%).

Out of the total 29,886 microbial barcodes, only 2,108 were also assigned to host cells, suggesting that the bulk of bacteria found in BAL samples exist as free-living organisms or in bacterial biofilms. Although microscopic evaluation would be needed to validate this hypothesis, bacterial biofilms have been previously documented in bronchoalveolar lavages[23], and the enrichment for host cells in these samples via centrifugation[18] may have also indirectly enriched these specimens for biofilm and/or host-associated microbes. However, for the fraction of bacteria associated to host cells, the distribution across disease types was not random. We found that while 2.3% of the non-COVID-19 patient cells were associated to bacterial cells, almost double (4%) could be observed in COVID-19 patients (Fig. 3a; Chi-squared test; p value = $3.75{\cdot}10^{-35}$). However, because COVID-19 diagnosis is highly correlated with mechanical ventilation in this cohort, this effect could be due to higher intubation rates in COVID-19 patients and possibly, a higher incidence of VAP. Within COVID-19 patients, we also evaluated the overlap between bacteria-associated host cells and cells with detected SARS-CoV-2 reads[18] (Supplementary Table 2). Out of 1,033 host cells associated with bacteria in COVID-19 patients and 343 cells with detected SARS-CoV-2 reads, only one cell was positive for both viral and bacterial reads. A binomial test for independence of virus and bacteria detection in the same host cell, showed that the observed co-occurrence in one cell only was highly unlikely (p value = $5.7{\times}10^{-4}$), therefore suggesting mutual exclusion of microbiome members and viruses in the same host immune cells. However, it must be noted that lack of detection does not necessarily imply lack of association of bacteria/virus to host cells, especially with experimental methods such as single-cell RNA-seq, designed for profiling host cells and not optimized for detection of these entities. Therefore, further studies with larger sample sizes are required to validate the co-exclusion hypothesis.

We also explored whether host-associated bacterial reads would preferentially be linked with specific cell types, taking into account the varying frequencies of cell types in COVID-19 patients and controls (see Methods). Such a preferential association would suggest that these observations are biologically relevant and not an artifact of the single-cell sample and library preparation. Among control patients, cell types were similarly distributed in both groups (i.e., with and without bacteria), with only a preferential association of microbial cells with neutrophils (p value = 3.61×10^{-12}; Fig. 3b; Supplementary Figure 4). However, in COVID-19 patients, three cell types were significantly associated with bacteria: neutrophils (p value = 8.27×10^{-29}), monocytes (p value = 4.82×10^{-5}) and monocyte-derived macrophages (p value = 2.23×10^{-51}; Fig. 3b; Supplementary Fig. 5 and Supplementary Data 3). We also found that different bacteria associate with distinct host cells. For instance, in COVID-19 patients, bacteria from the *Mycoplasma* genus preferentially associated to monocyte-derived macrophages (p value = 2.28×10^{-7}), while *Rothia* (p value = 8.21×10^{-4}), *Enterobacter* (p value = 2.59×10^{-5}), or *Klebsiella* (p value = 3.12×10^{-9}) are enriched in monocytes (Fig. 3c, Supplementary Data 3).

Last, we investigated whether the associations of bacteria to host cells are linked to host cell expression. To do so, we assessed whether expression-based cell subtype classification[18] for neutrophils, monocytes and macrophages showed non-random associations with bacteria across all samples in this cohort. Among the neutrophils, a subtype of inflammatory neutrophils characterized by expression of the calgranulin *S100A12* gene was enriched in bacteria-associated cells (Chi-squared two-sided test, p value 7.18×10^{-6}; Fig. 3d, e; Supplementary Data 3). This subset of cells was also found to be enriched in SARS-CoV-2 nucleocapsid gene reads[18], suggesting that the same cell type responsible for defense against the virus would be responding to potentially invasive bacteria in the lung. This subgroup is characterized by the expression of the calprotectin subunits encoded by *S100A8* and *S100A9*. It is known that S100A8/A9 heterodimer secretion is increased in infection-induced inflammation and has some antibacterial effects mediated by secretion of proinflammatory cytokines, release of reactive oxygen species and recruitment of other inflammatory cells, as well as chelation of Zn^{2+} necessary for bacterial enzymatic activity[24]. These mechanisms are mediated by binding of the S100A8/A9 dimer to TLR4 receptors to trigger the release of proinflammatory cytokines, such as IL-6 and TNF-α, and thus may contribute to sustain or exacerbate inflammation[25]. Therefore, the association with bacteria could, at least in part, explain the inflammatory phenotype of this neutrophil subset. To further examine this hypothesis, we explored differential gene expression between bacteria-associated and non-associated S100A12^{hi} neutrophils (Supplementary Data 4). Because association of these cells with SARS-CoV-2 and with bacteria was mutually exclusive, we also compared these changes with the ones triggered by the virus in neutrophils[18]. Within this subset, neutrophils with co-occurring bacteria showed significantly higher expression (Bonferroni-corrected p value < 0.05) of proinflammatory genes, including the cytokine *IL1B* and some of its target genes (*PTSG2*), the transcription factors encoded by *FOS* and *JUN*, and several genes encoding proteins involved in degranulation (*S100A9*, *FOLR3*, *HSPA1A*, *HSP90AA1*, *FCGR3B*), (Supplementary Data 4). Among these, *FOLR3*, a gene encoding for a folate receptor, is found in neutrophil secretory granules and has antibacterial functions, by binding folates and thus depriving bacteria of these essential metabolites[26]. This response differed to that of virus-engulfing neutrophils in that IFN response genes are not distinctively upregulated by bacteria.

Regarding myeloid cells, both inflammatory IL1B^{hi} monocytes (p value < 2×10^{-16}), as well as a mixed group of *CCL2*-expressing macrophages (p value = 5.38×10^{-10}) are enriched in bacteria-associated cells (Fig. 3f). These inflammatory monocytes are believed to have an important role in the aberrant immune response occurring in severe COVID-19 patients. In this case, further gene expression patterns were detected, specific for bacteria-associated cells: for CCL2^{hi} macrophages, cells with co-

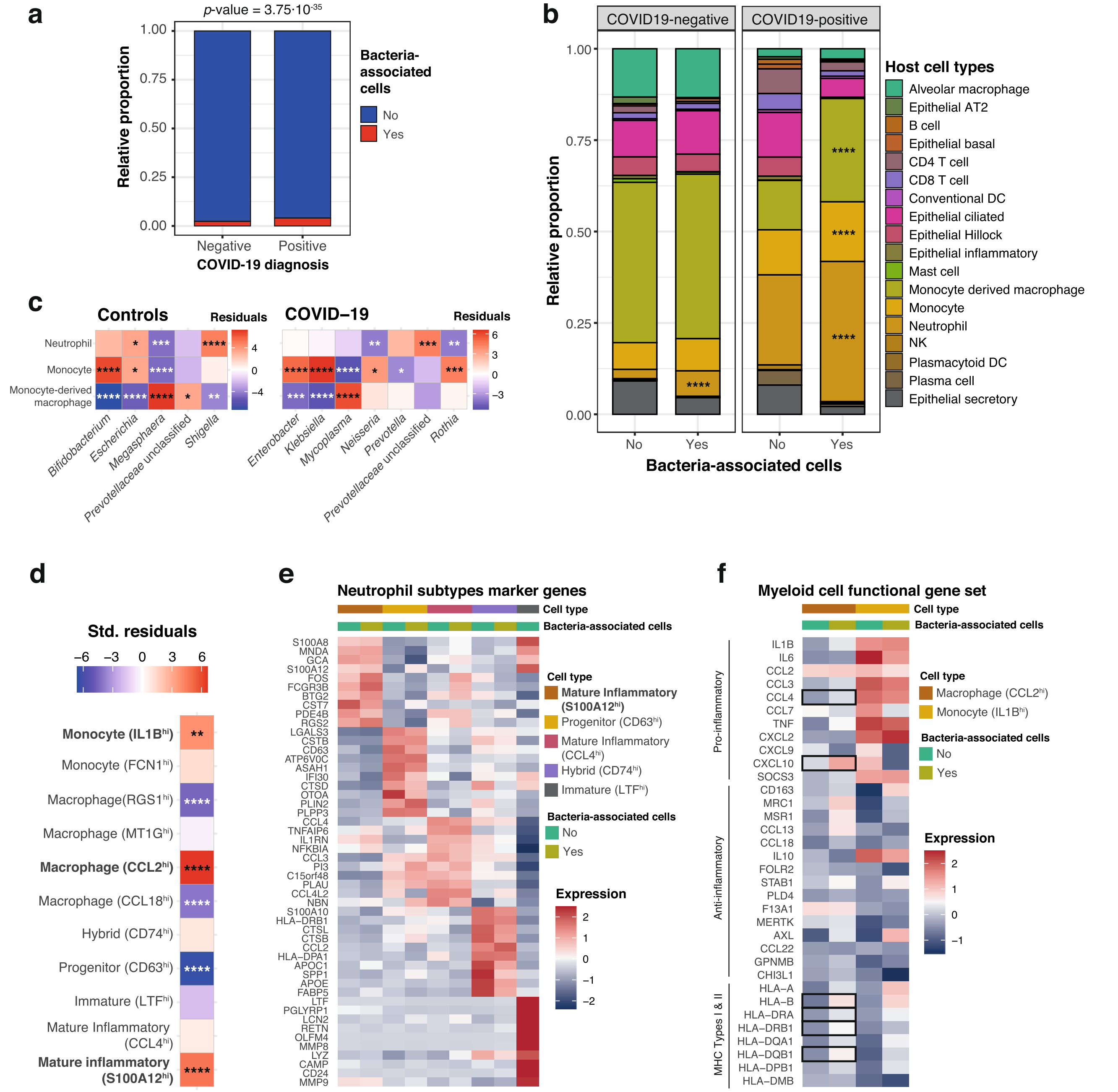

Fig. 3 Host single cells associated to the lower respiratory tract microbiota. a relative proportion of cells from negative ($n = 31,923$ cells) and positive ($n = 33,243$ cells) COVID-19 patients with (red) and without (blue) associated bacteria. The p value of a two-tailed Chi-squared test using the count data is shown on top. **b** Cell types enriched in bacteria-associated cells. Barplots represent the proportion of cell types without ("No"; $n = 63,463$ cells) and with ("Yes"; $n = 1,703$ cells with 2108 bacteria detected) bacteria in COVID-19 positive and negative patients. For each patient class, we tested for enrichment of bacteria-associated cells ("Yes") across the different cell types, using the proportions of non-bacteria associated cells ("No") as background (Chi-squared tests, two-tailed). **c** Bacterial genera preferentially associating to specific cell types. The heatmaps show the standardized residuals of a two-tailed Chi-squared test including all bacterial genera and the three host cell types enriched in bacteria, for controls (left) and COVID-19 positive patients (right). Taxa with no significant associations with any of the cell types are not shown. Enrichments are shown in red; depletions are depicted in blue. **d** Host cell subtypes associated with bacteria. The heatmap shows the standardized residuals of a two-tailed Chi-squared test, including the subtypes of neutrophils, monocytes and monocyte-derived macrophages with associated bacteria, considering cells without bacteria as background. Enrichments are shown in red; depletions are depicted in blue. **e** Marker genes detected for the 5 different subtypes of neutrophils, showing within-group differences between bacteria-associated and bacteria-non-associated cells. **f** Myeloid cell functional gene set showing the expression of canonical proinflammatory, anti-inflammatory and MHC-encoded genes for the two subtypes of myeloid cells significantly associated with bacteria (CCL2^{hi}-macrophages and IL1B^{hi}-monocytes). The heatmap also shows within group differences between bacteria-associated and non-associated cells. Statistically significant differences (two-sided Wilcoxon rank-sum test, Bonferroni-adjusted p value < 0.05) are marked with squares. For (**b–d**) asterisks denote significance after BH multiple testing correction as follows: *p value ≤ 0.05; **p value ≤ 0.01; ***p value ≤ 0.001; ****p value ≤ 0.0001; exact p values are provided in Supplementary Data 3. Source data are provided as a Source Data file.

occurring bacteria expressed higher levels of MHC genes of type I and II, suggesting a more active role of these cells in antigen presentation (Bonferroni-corrected p value < 0.05; Fig. 3f; Supplementary Data 4). A similar increase was also observed in monocytes, yet not significant (Supplementary Data 4), possibly due to the lower monocyte abundances in this dataset. Additionally, bacteria-associated macrophages express significantly higher levels of the calprotectin subunit genes *S100A8*/*A9*, similarly to neutrophils, as well as genes encoding proinflammatory chemokines (such as *CCL4*, *CXCL10*, and *CXCL1*).

Altogether, our results suggest that the bacteria detected in these cell subsets via scRNA-seq analyses may be contributing to the inflammatory response observed in the host.

Discussion

Since the beginning of the COVID-19 pandemic, a massive global effort by the scientific community was undertaken to understand physiopathology of SARS-CoV-2 infection and risk factors affecting disease outcome. In this study, we explored the respiratory microbiota as a potential risk factor for disease severity, and we evaluated the upper and lower respiratory tract microbiota in COVID-19 patients throughout hospitalization. We linked these data to viral load measurements and immunoprofiling results from nCounter and single-cell RNA sequencing data. To assess robustness of previously reported signals, we investigated the effect of potential confounders based on a broad panel of patient metadata variables.

We found that in the upper respiratory tract, SARS-CoV-2 viral load has a mild negative association with bacterial biodiversity. A larger effect of severity indicators such as calprotectin levels or length of ICU stay was observed, with diversity decreasing throughout the length of the ICU period, a pattern reminiscent of that seen in other pulmonary conditions[27,28]. The effect of ICU length-of-stay may be mediated by treatment options such as the administration of broad-spectrum antibiotics and/or patient intubation and mechanical ventilation. Antibiotic usage might also explain why calprotectin levels correlate with alpha diversity: such antibiotics would decrease overall microbial diversity, including (some of) the taxa that could be linked to inflammation. The observed effects of these clinical practices on microbiome alpha diversity could potentially explain why previous studies on the microbiota of COVID-19 patients show conflicting results regarding diversity: some studies reported lower diversity in sputum or throat swab samples of COVID-19 patients[14,15,17] while others, focusing on the lower respiratory microbiome using bronchoalveolar fluid samples, showed higher bacterial diversity in COVID-19 patients than in controls[16]. To further complicate matters, it cannot be excluded that sampling site or processing could also be potential confounders in these studies and/or reflect the different pathologies in the different areas of the respiratory tract.

We further found that between-patient microbiome variation (as measured by genus-level microbial beta-diversity) was also influenced by different severity indicators, such as the clinical status of the patient, or more importantly the type of oxygen support received, with mechanically ventilated patients harboring a different microbiota than nonintubated patients. This effect could not be fully explained by neither general antibiotic administration, nor the usage of specific antibiotics such as ceftriaxone, meropenem, or piperacillin-tazobactam, suggesting an independent effect of mechanical ventilation. Such an independent effect has previously been suggested in small cohorts[27,29,30], but it needs to be validated in larger studies. However, other associated practices, such as decontamination procedures, could still be responsible for the observed associations, and we cannot completely rule out a confounding effect of disease severity, as COVID-19 severity dictates the clinical practices executed. The impact of oxygen support was also reflected at the species- and strain-levels, with intubation causing a significant decrease and increase, respectively, in diversity. Combined, these results suggest that noninvasive ventilation (e.g., BIPAP, CPAP) can have microbial effects indicating that any form of ventilation may be a tipping point for microbial community differences.

Importantly, several of the taxa reported to change between intubated and nonintubated patients were reported to be linked to diagnosis in previous COVID-19 microbiome studies[14–16]. In our study, no taxa were specifically linked to SARS-CoV-2 viral load after controlling for mechanical ventilation. This result suggests the possibility that mechanical ventilation and its associated clinical practices are confounding previous results. Indeed, one study comparing COVID-19 patients with patients diagnosed of community-acquired pneumonia found no differences in respiratory microbiome composition between both groups of patients, but both groups did differ from healthy controls[31]. Together, these results indicate that patient intubation or even noninvasive ventilation, as well as their associated medical practices, are to be considered as important confounders when studying the upper respiratory microbiome, and we strongly suggest future COVID-19 microbiome studies should foresee and include strategies to account for this covariate. As an example, a recent study found a single ASV corresponding to the genus *Rothia* that was specific for SARS-CoV-2 patients after controlling for ICU-related confounders by comparing with a previous study of the microbiome in ICU patients[32]. Additionally, these findings on potential drivers of microbiome variation are not exclusive to COVID-19 disease: the effect of intubation on the respiratory microbiome and its influence on the incidence of ventilator-associated pneumonia have been previously studied[27,29,30].

To better understand the potential functional consequences of these procedures and linked microbial shifts, we also profiled the microbiome of the lower respiratory tract using single-cell data obtained from a cross-sectional cohort of patients derived from the same hospital. Our results show that single-cell RNA-seq, despite not being optimized for microbial detection and profiling, can identify bacteria alone or in association with specific human cells. In this cohort, we identified a subset of bacteria associated with host cells, more specifically with neutrophils, monocytes and macrophages. While associations with neutrophils were common to COVID-19 patients and controls, associations with monocytes and macrophages were unique to COVID-19 cells. This enrichment shows that these bacteria are likely not random contaminants, from which an even distribution across cell types (i.e., considering cell type abundances) would be expected. The identity of these host cells suggests that bacteria could have been phagocyted by these innate immune system cells, rather than be attached to the host cell surface. To the best of our knowledge, this is the first study linking interacting host cells and lung microbiome via high-throughput single-cell RNA-seq.

We find that host cells associated with bacteria, most of which are of oral origin, exhibit proinflammatory phenotypes as well as higher levels of MHC for antigen presentation. In this single-cell cohort it was observed that critical COVID-19 patients are characterized by an impaired monocyte to macrophage differentiation, resulting in an excess of proinflammatory monocytes, as well as by prolonged neutrophil inflammation[18]. Given that only these cell types are enriched in bacteria, we hypothesize that the respiratory (or ventilation-linked) microbiome may be playing a role in exacerbating COVID-19 progression in the lower respiratory tract. We verified that this response could likely be driven by bacteria and not SARS-CoV-2, which is also detected mostly in these cell types, as there is almost no overlap in detection of both the virus and bacteria in the same cells. However, it must be noted that lack of detection does not completely rule out presence of both virus and bacteria within these cells. Therefore, further research is required in order to confirm a causative role of the microbiota in this immune impairment characteristic of critical disease, and to reveal the specific mechanisms involved.

The presence of oral taxa in the lung microbiota is not unique of disease conditions. It is known that microaspiration, or the aspiration of aerosol droplets originated in the oral cavity, occurs in healthy individuals and can serve as a route for lung colonization of oral commensals[33]. Such an increase of oral bacteria in the lower respiratory tract could be facilitated when critically ill patients—including but not limited to COVID-19—get intubated. Indeed, oral bacteria have been linked to ventilator-associated pneumonia[34,35]. It is yet to be elucidated whether COVID-19 physiopathology favors lung colonization by oral bacteria or if, in contrast, a lung microbiome previously colonized by oral microbes could also contribute to the disease. What is known is that an increase of oral bacteria in the lower respiratory tract can result in an increased inflammatory phenotype, even in healthy subjects[36]

Overall, this study provides a systematic analysis of potential confounders in COVID-19 microbiome studies. We identified that ICU hospitalization and type of oxygen support, which may be at least partially explained by clinical practices such as antibiotic usage, had large impacts on the upper respiratory tract microbiome and have the potential to confound clinical microbiome studies. Among the different types of oxygen support we reported the largest shifts in microbial community structure between the intubated and non-intubated patients. We found that oral taxa were strongly enriched in the upper respiratory tract of mechanically ventilated COVID-19 patients, and specific taxa were also found in the lower respiratory tract of COVID-19 patients. Further, in the lower respiratory tract, microbes were strongly associated with specific proinflammatory immune cells. This information contributes to a collective body of literature on the pathology of COVID-19 and suggests that careful attention be paid to ICU stay and type of oxygen support and associated clinical practices such as antibiotic usage or oral decontamination procedures when evaluating the role of the lung microbiome on COVID-19 disease progression.

Methods

Study design and patient cohorts. All experimental protocols and data analyses were approved by the Ethics Commission from the UZ Leuven Hospital, under the COntAGIouS observational clinical trial (study number S63881). The study design is compliant with all relevant ethical regulations, including the Declaration of Helsinki and the GDPR. All participants gave their informed consent to participate in the study. Research was performed as part of the COntAGIouS observational clinical trial, as recorded in: https://clinicaltrials.gov/ct2/show/NCT04327570.

A total of 58 patients from the COntAGIouS observational trial were included as our upper respiratory tract cohort. All patients were admitted to the UZ Leuven hospital with a diagnostic of COVID-19. The disease was diagnosed based on a) a positive qRT-PCR test, performed on admission or previously on other hospitals, when patients were transferred from other medical facilities; or b) a chest CT scan showing alveolar damage and clinical symptoms of the disease. All patients included in the study were admitted to ICU for a variable amount of time. Nasopharyngeal swabs were taken from these patients at different timepoints throughout ICU stay and after ICU discharge, during recovery in ward. A total of 112 swabs were processed for upper respiratory microbiome characterization (Fig. 1a).

To extend our findings from the upper respiratory tract, we also profiled the lower respiratory tract microbiota in a different cohort[18] of 35 patients. These patients were recruited as part of the same observational trial at UZ Leuven hospital, but there was no overlap between these 35 patients and the 58 patients from the upper respiratory tract cohort. The lower respiratory tract cross-sectional cohort is composed by 22 COVID-19 patients and 13 pneumonitis controls with negative qRT-PCR for SARS-CoV-2, with varying disease severity. Previous data from single-cell RNA-sequencing had been collected for this cohort[18]. We reanalyzed this single-cell dataset to profile the lower respiratory tract microbiota in these patients.

RNA/DNA extraction and sequencing. Nucleic acid extraction from the swab samples was performed with AllPrep DNA/RNA/miRNA Universal kit (QIAGEN, catnr. 80224). Briefly, swabs from the potentially infectious samples were inactivated by adding 600 µL RLT-plus lysis buffer. To increase bacterial cell lysis efficiency, glass beads and DX reagent (Pathogen Lysis Tubes, QIAGEN, catnr. 19091) were added to the lysis buffer, and samples were disrupted in a FastPrep-24^{TM} instrument with the following program: 1-minute beating at 6.5 m/sec, 1-minute incubation at 4 °C, 1-minute beating at 6.5 m/sec, 1-minute incubation at 4 °C. After lysis, the remaining extraction steps followed the recommended protocol from the manufacturer. DNA was eluted in 50 µL EB buffer. Amplification of the V4 region of the 16S gene was done with primers 515F and 806R (GTGYCAG CMGCCGCGGTAA/ GGACTACNVGGGTWTCTAAT), using single multiplex identifiers and adaptors as previously described[37]. RNA was eluted in 30 µL of nuclease-free water and used for SARS-CoV-2 viral load determination in the swabs as well as to measure inflammatory markers and cytokines and to estimate host cell populations via marker gene expression using nCounter. In brief, raw nCounter data were processed using nSolver 4.0 software (Nanostring), sequentially correcting three factors for each individual sample: technical variation between samples (using spiked positive control RNA), background correction (using spiked negative control RNA) and RNA content variation (using 15 housekeeping genes). We have previously validated nCounter digital transcriptomics for simultaneous quantification of host immune and viral transcripts[38], including respiratory viruses in nasopharyngeal aspirates, even with low RNA yield[39–41].

DNA sequencing was performed on an Illumina MiSeq instrument, generating paired-end reads of 250 base pairs. For quality control, reads were demultiplexed with LotuS v1.565[42] and processed following the DADA2 microbiome pipeline using the R packages DADA2[43] and phyloseq[44]. Briefly, reads were filtered and trimmed using the parameters truncQ=11, truncLen=c(130,200), and trimLeft=c(30, 30) and then denoised. After removing chimeras, amplicon sequence variants (ASVs) table was constructed and taxonomy was assigned using the Ribosomal Database Project (RDP) classifier implemented in DADA2 (RDP trainset 16/release 11.5). The abundance table was then corrected for copy number, rarefied to even sequencing depth, and decontaminated. For decontamination, we used the prevalence-based contaminant identification method in the R package decontam[45].

16S statistical analysis. All the 16S data analyses were performed using R v3.6.0 and the packages vegan (v2.5.7)[46], phyloseq (v1.34.0)[44], CoDaSeq (v0.99.6)[47], DESeq2 (v1.30.1)[48], Biostrings (v2.58.0)[49], rstatix (v0.7.0)[50], glmulti (v1.0.8)[51], sjPlot (v2.8.7)[52], and DECIPHER (v2.18.1)[53].

To analyze the 16S amplicon data, technical replicates were pooled and counts from technical replicates were added. For all the analyses using genus-level agglomerated data, only samples containing more than 10,000 reads assigned at the genus level were used (101 samples in total). Alpha-diversity was analyzed using Shannon's Diversity Index. Comparison of the alpha diversity values across different groups was performed using Kruskal–Wallis tests for comparisons across multiple groups. When applicable, pairwise comparisons were performed using Dunn post hoc tests. To establish the potential associations of alpha diversity with different metadata variables, we selected eight variables related to COVID-19 disease and/or known to affect microbiome composition and diversity: patient ID, days spent in ICU, SARS-CoV-2 viral load, antibiotic usage for ceftriaxone and meropenem/piperacillin-tazobactam, previous mechanical ventilation, calprotectin gene expression and CRP levels. Meropenem and piperacillin-tazobactam were merged as a single antibiotic as their administration is indicated under the same clinical guidelines.

We used the R package glmulti to perform an exhaustive evaluation of the 256 models including all possible combinations of the selected variables. All models generated were generalized linear models or generalized linear mixed models (when including the patient ID as a random effect), using a Gaussian family with a logarithmic link. Model ranking and selection was performed based on the lowest small-sample-corrected Akaike Information Criterion (AICc), and model significance was assessed comparing with a null model (including the intercept only) using the ANOVA test. The final variable importance was calculated as a weighted average of the models in which each of the variables appeared, with weights corresponding to the model ranks, defined by their AICc values. This was also implemented as part of the glmulti package. The final model plots were generated with the sjPlot package. Intra-patient differences in alpha diversity between timepoints before and after administration of antibiotics or mechanical ventilation were determined with Wilcoxon signed-rank tests.

Beta diversity analyses were performed using distance-based redundancy analyses (dbRDA), using Aitchison distances. Prior to CLR data transformation, we filtered the data using the CoDaSeq.filter function, to keep samples with more than 10,000 reads and taxa with a relative abundance above 0.1% in any sample, as well as a prevalence of at least 10% in the cohort. To replace zeros, we first calculated the minimum (nonzero) relative abundance of each taxon across all samples. Then, for samples with zero counts for a given taxon, the minimum relative abundance of the specific taxon was multiplied by the total counts of such samples and this value was used to impute the zeros. dbRDA analyses were performed using the capscale function from vegan, first in univariate analyses with 72 metadata variables

(Supplementary Data 1). Model p values were corrected using Benjamini-Hochberg's (BH) multiple-testing correction, to select 20 variables with BH-adjusted$p$ values < 0.05. These 20 variables were included in a multivariate model, and nonredundant contribution to variation was calculated using forward stepwise variable selection via the ordiR2step function from vegan. To deconfound the effect of antibiotics and patient ID for oxygen support type, partial dbRDA was used, including both the antibiotics and patient ID as blocking variables. Metadata variables containing dates, as well as noninformative metadata were excluded. Noninformative metadata variables were defined as those containing a single non-NA value or, for categorical variables, those being unevenly distributed (with $> 90\%$ of the samples belonging to the same category, for instance, an antibiotic administered only in two different samples). Additionally, from pairs of highly collinear variables (correlation higher than 0.9), only one variable was kept.

Differential taxa abundance analyses were performed using DESeq2's likelihood ratio tests and controlling for potential confounders when indicated, including them in the null model.

To explore species-level and strain-level diversity, 16S sequences were first clustered into 97% nucleotide diversity operational taxonomic units (OTUs) using the *R* packages Biostrings and DECIPHER. These OTUs were used to represents the species-level. The number of unique 16S sequences clustered within each OTU were used to represent the number of detectable strains per species. To calculate strain-level diversity per sample, the number of strains of 5 detected OTU species were randomly selected and averaged. This was repeated 1000× and the average of the all 1000 subsamplings was used as the final strain-level diversity value for each sample, as previously described[54]. To account for uneven sampling assessing diversity differences based on different parameters, we randomly selected and averaged the species- and strain-level diversity of five samples per parameter. This was repeated 100× and the subsamplings were used to assess the significant differences between species- and strain-level diversity across the parameters. The average of all 100 subsamplings was used to as the input for a Pearson's correlation between species- and strain-level diversity.

All statistical tests were two-sided unless otherwise specified, and when multiple tests were applied to the different features (e.g., the differential taxa abundances across two conditions) p values were corrected for multiple testing using Benjamini–Hochberg's method.

Identification of microbial reads in BAL scRNA-seq data. BAL scRNA-seq raw fastq data, as well as cell type and subtype assignations for all individual cells, were obtained from a previous publication from within the COntAGIouS consortium. Experimental procedures on BAL samples as well as detailed host single-cell gene expression analyses are detailed in the original publication[18].

5' single-cell RNA-seq data obtained from the 10X Genomics Chromium platform was processed with an in-house pipeline to identify microbial reads. This pipeline comprises a series of steps designed to detect bacterial reads with high sensitivity, while discarding potential false positives. For microbial identification, only the read 2 fastq file from the raw sequencing files, containing the information on the cDNA fragment, was used. Trimmomatic[55] (v0.38) was used to trim low quality bases and adapters, and discard short reads. Additionally, Prinseq + +[56] (v1.2) was used to discard reads with low-complexity stretches such as poly-A sequences. Following these two quality control steps, reads from human and potential sequencing artifacts (phage phiX174) were mapped with STAR[57] (v2.7.1) and discarded. The remaining unmapped reads were mapped against reference microbial genomes using a two-step approach: first, we scanned these remaining reads using mash screen[58] (v2.0) against a custom database of 11,685 microbial reference genomes including bacteria, archaea, fungi and viruses. Genomes likely to be present in the analyzed sample (selected using a threshold of at least two shared hashes from mash screen) were selected and reads were pseudoaligned to this subset of reference genomes using kallisto[59] (v0.44.0). Kallisto provides two outputs: an "abundances" table containing the number of reads aligned to each gene from the pre-selected set of reference genomes and a pseudoalignment file (in *.bam format) containing the mapping information for each of the reads processed by kallisto. From the abundances table, we derived a taxonomy table, assigning each gene to its corresponding species, as well as a functional table, mapping each gene to KEGG functional annotation using KEGG Orthology numbers (KOs). To remove potential artifacts, two additional filters were applied to the taxonomic table: first, if less than 10 different functions (i.e., 10 different KOs) were expressed from a given species, such species was discarded. This filter ensures identification of active bacteria, minimizing the capture of contaminants appearing during the sample preparation or sequencing. Second, if one function accounted for more than 90% of the mapped reads of a given species, it was also discarded. This filter was aimed at removing potential artifacts caused by errors in the reference genome assemblies from our database.

Bacterial reads were assigned their specific barcodes and UMIs as follows: read IDs from the mapped microbial reads were retrieved from the kallisto pseudoalignment (*.bam) output using SAMtools (v1.9)[60]. These unique read IDs were used to retrieve the specific barcodes and UMIs using the raw read 1 fastq files, thus assigning each barcode and UMI univocally to a microbial species and function. Barcodes assigned to bacterial species that had been removed in the last two filtering steps of the single-cell analysis pipeline (see above) were discarded, to avoid including potential contaminants in the host-bacteria association analyses.

Direct associations between bacteria and host cells. Host single-cell transcriptomics data was obtained from the Seurat[61] object after preprocessing and integrating the samples of the single-cell cohort, as described previously[18]. From the Seurat object, the metadata was extracted, including the information on patient group (COVID-19 or control) and severity of the disease (moderate or critical) as well as cell type and subtype annotation corresponding to each barcode. Enrichment of bacteria detected in patient groups or cell types was calculated using Chi-squared tests, with effect sizes determined via the standardized residuals. Significance was assessed via post hoc tests using the R package chisq.posthoc.test[62].

To evaluate the overlap between bacterial and viral reads detection in host cells of COVID-19 patients, we considered the total number of cells analyzed in these patients: 33,243. Of these, 31,868 cells do not have associated bacterial or viral reads; 1032 have only bacterial reads; 342 have only viral reads; and 1 has both viral and bacterial reads detected (Supplementary Table 2). The marginal probability for bacterial detection is thus P(bacterial detection) $= 1033/33{,}243 = 0.031$; while the marginal probability for viral detection in this dataset is P(viral detection) $= 343/33{,}243 = 0.010$. Assuming independence of both events, the joint probability of finding a host cell associated to both bacterial and viral reads would be P(bacterial and viral detection) $= 0.031{*}0.010 = 3.2 \times 10^{-4}$. With this joint probability and a total of 33,243 cells profiled, an average of 10.65 host cells should have both bacterial and viral reads detected. A Chi-squared test suggests nonindependence of the data (p value $= 4.1 \times 10^{-3}$). Additionally, we performed an exact binomial test, considering number of successes $= 1$ (joint bacterial and viral detection), probability of success $= 3.2 - 10^{-4}$, (the joint probability assuming independence of both events), and number of trials $= 33{,}243$ (the total number of cells studied). This two-sided test evaluates the null hypothesis that the joint probability of both events is the one calculated assuming independence. The result of this test (p value $= 5.7 \times 10^{-4}$) suggests rejecting the null hypothesis. Therefore, these analyses altogether suggest that both events are not independent and that there is mutual exclusion of microbiome members and viruses in the same host immune cells.

For cell types showing an enrichment in associated bacteria, a new Seurat object was created by subsetting the specific cell type. Chi-squared tests were also used to determine the enrichment of bacteria-associated cell subtypes. Previous annotations of cell subtypes[18] were used to generate new clusters manually and identify marker genes for these subtypes, using the function findAllMarkers from Seurat. This function was also used to find differentially expressed genes between bacteria-associated and not-bacteria-associated host cells of each subtype. When using this function, reported adjusted p values are calculated using Bonferroni correction by default.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw amplicon sequencing data generated in this study, as well as patient metadata, have been deposited in the European Genome-phenome Archive (EGA) repository under accession code EGAS00001004951. These data are available under restricted access to comply with current European personal data protection regulations. Access can be obtained for research purposes within specific approved projects granted by the corresponding Data Access Committee (raw sequencing data requests should be addressed to Prof. Jeroen Raes, metadata requests should be addressed to Prof. Joost Wauters: https://ega-archive.org/dacs/EGAC00001001901). The single-cell RNA-seq data was first described in a separate publication[18] and deposited also in EGA with accession number EGAS00001004717. Source data are provided with this paper.

Code availability

The computer code used to process the data generated for this study has been made available at https://raeslab.org/software/COVID19_respiratory_microbiome. A release version of the code has been archived in the Zenodo repository: https://doi.org/10.5281/zenodo.5524635 (https://zenodo.org/record/5524635)[63]. Source data are provided with this paper.

Received: 16 August 2021; Accepted: 6 October 2021;
Published online: 29 October 2021

References

1. Zhou, F. et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395**, 1054–1062 (2020).
2. Grasselli, G. et al. Risk Factors Associated with Mortality among Patients with COVID-19 in Intensive Care Units in Lombardy, Italy. *JAMA Intern. Med.* **180**, 1345–1355 (2020).
3. Mikami, T. et al. Risk factors for mortality in patients with COVID-19 in New York City. *J. Gen. Intern. Med.* **36**, 17–261 (2020).

4. Guo, W. et al. Diabetes is a risk factor for the progression and prognosis of COVID-19. *Diabetes. Metab. Res. Rev.* **36**, e3319 (2020).
5. Lighter, J. et al. Obesity in Patients Younger Than 60 Years Is a Risk Factor for COVID-19 Hospital Admission. *Clin. Infect. Dis.* **71**, 896–897 (2020).
6. Yu, X. et al. SARS-CoV-2 viral load in sputum correlates with risk of COVID-19 progression. *Crit. care* **24**, 170 (2020).
7. Magleby, R. et al. Impact of Severe Acute Respiratory Syndrome Coronavirus 2 Viral Load on Risk of Intubation and Mortality Among Hospitalized Patients With Coronavirus Disease 2019. *Clin. Infect. Dis.* **30**, ciaa851 (2020).
8. Westblade, L. F. et al. SARS-CoV-2 Viral Load Predicts Mortality in Patients with and without Cancer Who Are Hospitalized with COVID-19. *Cancer Cell.* **38**, 661–671.e2 (2020).
9. Leisman, D. E. et al. Cytokine elevation in severe and critical COVID-19: a rapid systematic review, meta-analysis, and comparison with other inflammatory syndromes. *Lancet Respiratory Med.* **8**, 1233–1244 (2020).
10. Sinha, P., Matthay, M. A. & Calfee, C. S. Is a 'cytokine Storm' Relevant to COVID-19? *JAMA Intern. Med.* **180**, 1152–1154 (2020).
11. Khatiwada, S. & Subedi, A. Lung microbiome and coronavirus disease 2019 (COVID-19): Possible link and implications. *Hum. Microbiome J.* **17**, 100073 (2020).
12. Dickson, R. P., Martinez, F. J. & Huffnagle, G. B. The role of the microbiome in exacerbations of chronic lung diseases. *Lancet* **384**, 691–702 (2014).
13. Huffnagle, G. B., Dickson, R. P. & Lukacs, N. W. The respiratory tract microbiome and lung inflammation: A two-way street. *Mucosal Immunol.* **10**, 299–306 (2017).
14. Zhang, H. et al. Metatranscriptomic Characterization of COVID-19 Identified A Host Transcriptional Classifier Associated With Immune Signaling. *Clin. Infect. Dis.* **73**, 376–385 (2020).
15. Xu, R. et al. Temporal dynamics of human respiratory and gut microbiomes during the course of COVID. *medRxiv.* 2020.07.21.20158758 (2020).
16. Han, Y., Jia, Z., Shi, J., Wang, W. & He, K. The active lung microbiota landscape of COVID-19 patients. *medRxiv* 2020.08.20.20144014 (2020).
17. Mostafa, H. H. et al. Metagenomic Next-Generation Sequencing of Nasopharyngeal Specimens Collected from Confirmed and Suspect COVID-19 Patients Downloaded from. *MBio* **11**, e01969–20 (2020).
18. Wauters, E. et al. Discriminating mild from critical COVID-19 by innate and adaptive immune single-cell profiling of bronchoalveolar lavages. *Cell Res* **31**, 272–290 (2021).
19. Ho Man, W., de Steenhuijsen Piters, W. A. & Bogaert, D. The microbiota of the respiratory tract: gatekeeper to respiratory health. *Nat. Rev. Microbiol.* **15**, 259–270 (2017).
20. Whelan, F. J. et al. Longitudinal sampling of the lung microbiota in individuals with cystic fibrosis. *PLoS One* **12**, e017281 (2017).
21. Nolan, T. J. et al. Low-pathogenicity Mycoplasma spp. alter human monocyte and macrophage function and are highly prevalent among patients with ventilator-acquired pneumonia. *Thorax* **71**, 594–600 (2016).
22. Harris, J. R., Balajee, S. A. & Park, B. J. Pneumocystis jirovecii pneumonia: Current knowledge and outstanding public health issues. *Curr. Fungal Infect. Rep.* **4**, 229–237 (2010).
23. Dickson, R. P. et al. Cell-associated bacteria in the human lung microbiome. *Microbiome* **2**, 28 (2014).
24. Wang, S. et al. S100A8/A9 in inflammation. *Front. Immunol.* **9**, 1298 (2018).
25. Coveney, A. P. et al. Myeloid-related protein 8 induces self-tolerance and cross-tolerance to bacterial infection via TLR4- and TLR2-mediated signal pathways. *Nat. Publ. Gr.* https://doi.org/10.1038/srep13694 (2015).
26. Holm, J. & Hansen, S. I. Characterization of soluble folate receptors (folate binding proteins) in humans. Biological roles and clinical potentials in infection and malignancy. *Biochimica et. Biophysica Acta - Proteins Proteom.* **1868**, 140466 (2020).
27. Zakharkina, T. et al. The dynamics of the pulmonary microbiome during mechanical ventilation in the intensive care unit and the association with occurrence of pneumonia. *Thorax* **72**, 803–810 (2017).
28. Schmitt, F. C. F. et al. Pulmonary microbiome patterns correlate with the course of disease in patients with sepsis-induced ARDS following major abdominal surgery. *J. Hosp. Infect.* **24**, S0195–6701(20)30203-6 (2020).
29. Kelly, B. J. et al. Composition and dynamics of the respiratory tract microbiome in intubated patients. *Microbiome* **4**, 7 (2016).
30. Otsuji, K. et al. Dynamics of microbiota during mechanical ventilation in aspiration pneumonia. *BMC Pulm. Med.* **19**, 260 (2019).
31. Shen, Z. et al. Genomic Diversity of Severe Acute Respiratory Syndrome-Coronavirus 2 in Patients With Coronavirus Disease 2019. *Clin. Infect. Dis.: Off. Publ. Infect. Dis. Soc. Am.* **71**, 713–720 (2020).
32. Marotz, C. et al. Title: Microbial context predicts SARS-CoV-2 prevalence in patients and the hospital built environment. *medRxiv* 2020.11.19.20234229 (2020).
33. Bassis, C. M. et al. Analysis of the upper respiratory tract microbiotas as the source of the lung and gastric microbiotas in healthy individuals. *MBio* **6**, e00037 (2015).
34. Brennan, M. T. et al. The role of oral microbial colonization in ventilator-associated pneumonia. *Oral Surgery. Oral. Med. Oral. Pathol. Oral. Radiol. Endodontology* **98**, 665–672 (2004).
35. Stonecypher, K. Ventilator-associated pneumonia: the importance of oral care in intubated adults. *Crit. Care Nurs. Q* **33**, 339–347 (2010).
36. Segal, L. N. et al. Enrichment of the lung microbiome with oral taxa is associated with lung inflammation of a Th17 phenotype. *Nat. Microbiol.* **1**, 1–11 (2016).
37. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112–5120 (2013).
38. Moens, B. et al. Simultaneous RNA quantification of human and retroviral genomes reveals intact interferon signaling in HTLV-1-infected CD4+ T cell lines. *Virol. J.* **9**, 171 (2012).
39. Fukutani, K. F. et al. Pathogen transcriptional profile in nasopharyngeal aspirates of children with acute respiratory tract infection. *J. Clin. Virol.* **69**, 190–196 (2015).
40. Bouzas, M. L. et al. Diagnostic accuracy of digital RNA quantification versus real-time PCR for the detection of respiratory syncytial virus in nasopharyngeal aspirates from children with acute respiratory infection. *J. Clin. Virol.* **106**, 34–40 (2018).
41. Fukutani, K. F. et al. In situ immune signatures and microbial load at the nasopharyngeal interface in children with acute respiratory infection. *Front. Microbiol.* **9**, 2475 (2018).
42. Hildebrand, F., Tadeo, R., Voigt, A. Y., Bork, P. & Raes, J. LotuS: an efficient and user-friendly OTU processing pipeline. *Microbiome* **2**, 30 (2014).
43. Callahan, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
44. McMurdie, P. J. & Holmes, S. phyloseq: an R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* **8**, e61217 (2013).
45. Davis, N. M., Proctor, Di. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226 (2018).
46. Oksanen, J. et al. vegan: Community Ecology Package. (2019).
47. Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V. & Egozcue, J. J. *It's all relative: analyzing microbiome data as compositions. Ann. Epidemiol.* **26**, 322–329 (2016).
48. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
49. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: Efficient manipulation of biological strings. https://bioconductor.org/packages/Biostrings (2019).
50. Kassambara, A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. (2020).
51. Calcagno, V. & de Mazancourt, C. glmulti: An R package for easy automated model selection with (generalized) linear models. *J. Stat. Softw.* **34**, 29 (2010).
52. Lüdecke, D. sjPlot: Data Visualization for Statistics in Social Science. (2021).
53. Wright, E. S. Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *R Journal.* **8**, 352–359 (2016).
54. Gregory, A. C. et al. Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177**, 1109–1123.e14 (2019).
55. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
56. Cantu, V. A., Sadural, J. & Edwards, R. PRINSEQ + +, a multi-threaded tool for fast 1 and efficient quality control and 2 preprocessing of sequencing datasets. https://doi.org/10.7287/peerj.preprints.27553v1 (2019).
57. Dobin, A. et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
58. Ondov, B. D. et al. Mash Screen: high-throughput sequence containment estimation for genome discovery. *Genome Biol.* **20**, 232 (2019).
59. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
60. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
61. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
62. Ebbert, D. chisq.posthoc.test: A Post Hoc Analysis for Pearson's Chi-Squared Test for Count Data. http://chisq-posthoc-test.ebbert.nrw/ (2019).
63. Lloréns-Rico, V. et al. Code accompanying manuscript 'Clinical practices underlie COVID-19 patient respiratory microbiome composition and its interactions with the host'. https://doi.org/10.5281/ZENODO.5524635 (2021).

Acknowledgements

This study has been supported by funding from the VIB Grand Challenges Program. VLR is supported by an FWO senior postdoctoral fellowship (12V9421N). ACG is supported by an EMBO postdoctoral fellowship (ALTF 349-2019). The Raes lab is supported by KU Leuven, the Rega institute and VIB.

Author contributions

V.L.R., A.C.G., J.W., J.R. designed the study. S.J., T.V.W., J.N., C.D., J.G., G.H., P.M. collected and processed the BAL samples. P.V.M. and L.V. collected the clinical data. JW and EW collected the swabs. V.L.R., A.C.G. and J.V.W. processed the swabs. J.V.W., M.B., and S.M.M. determined SARS-CoV-2 viral loads and host gene expression from swabs. D.L. and J.Q. generated the single-cell raw data as well as the processed gene-count matrix with annotations of cell types and subtypes. V.L.R. and A.G.C. analyzed the data. V.L.R., A.C.G. and J.R. wrote the manuscript. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-26500-8.

Correspondence and requests for materials should be addressed to Jeroen Raes.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

CONTAGIOUS collaborators

Yannick Van Herck[14], Alexander Wilmer[13], Michael Casaer[12], Stephen Rex[15], Nathalie Lorent[9], Jonas Yserbyt[9,11], Dries Testelmans[9,11] & Karin Thevissen[16]

[14]Laboratory of Experimental Oncology, Department of Oncology, KU Leuven, Leuven, Belgium. [15]Anesthesiology and Algology, Department of Cardiovascular Sciences, KU Leuven, Leuven, Belgium. [16]Centre of Microbial and Plant Genetics, Department of Microbial and Molecular Systems (MS), KU Leuven, Leuven, Belgium.