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

Verónica Lloréns-Rico1,2, Ann C. Gregory1,2, Johan Van Weyenbergh3, Sander Jansen4, Tina Van Buyten4, Junbin Qian5,6, Marcos Braz3, Soraya Maria Menezes3, Pierre Van Mol7,8,9, Lore Vanderbeke10, Christophe Dooms9,11, Jan Gunst12, Greet Hermans12, Philippe Meersseman13, CONTAGIOUS collaborators*, Els Wauters9,11, Johan Neyts4, Diether Lambrechts7,8, Joost Wauters13,17 & Jeroen Raes1,2,17

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.
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 death1–3, as well as type 2 diabetes and obesity4–6. Other potential risk factors for critical condition and death are a viral load of the patient upon hospital admission6–8 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 failure9,10.

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 marker11. Previous studies exploring other pulmonary disorders have shown that the lung microbiota members may exacerbate symptoms and contribute to their severity12, potentially through direct crosstalk with the immune system and/or due to bacteremia and secondary infections13. First studies of the respiratory microbiome in COVID-19 have revealed elevated levels of opportunistic pathogenic bacteria14–16. However, reports on bacterial diversity are contradictory. While some studies report a low microbial diversity in COVID-19 patients14,17 that rebounds following recovery15, others show an increased diversity in the COVID-19 associated microbiota16. 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 data18 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 (Table 1). The upper respiratory samples, RNA were 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 nasopharynx19.

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 (22%) | 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) |


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 ($\text{AICc} = 121.79; p$ 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 1c).

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 and 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...
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 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 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 patients 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 in premechanical ventilation, 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 Megaspheara (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
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 conversely, 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, and the enrichment for host cells in these samples via centrifugation 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%) not random. We found that while 2.3% of the non-COVID-19 associated to host cells, the distribution across disease types was both viral and bacterial reads. A binomial test for independence was performed between bacteria-associated host cells and cells with detected bacteria-associated cells (Fig.3f). These in,

Regarding myeloid cells, both inflammatory IL1Bhi monocytes (p value < 2 x 10^{-16}), as well as a mixed group of CC2L2-expressing macrophages (p value = 5.38 x 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 CC2L2 macrophages, cells with co-

As a result, we observed that 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 x 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 x 10^{-20}), monocytes (p value = 4.82 x 10^{-5}) and monocyte-derived macrophages (p value = 2.23 x 10^{-31}; 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 x 10^{-7}), while Rothia (p value = 8.21 x 10^{-4}), Enterobacter (p value = 2.59 x 10^{-3}), or Klebsiella (p value = 3.12 x 10^{-3}) are enriched in monocytes (Fig. 3c; Supplementary Data 3).
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 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. 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 (CCL2hi-macrophages and IL18hi-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.
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).

Together, 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. 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, while others, focusing on the lower respiratory microbiome using bronchoalveolar fluid samples, showed higher bacterial diversity in COVID-19 patients than in controls.

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, 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. 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.

To better understand the potential functional consequences of these microbe-related microbiota 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 phagocytosed 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. 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. What 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 CONGALou observational clinical trial (study number 563881). 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 CONGALou’s observational clinical trial, as recorded in: clinicaltrials.gov/ct2/show/NCT034327570.

A total of 58 patients from the CONGALou observational trial were included as our upper respiratory tract cohort. All patients were admitted to the UZ Leuven hospital with a diagnosis of COVID-19. The disease was diagnosed based on a) a positive COVID-19 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 [37] 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 [38]. 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 infected area were subjected to nucleic acid extraction by adding 600 µL RLT-plus lysis buffer. To increase bacterial cell lysis efficiency, glass beads and DX reagent (Pathogen Lysis Tubes, QiAGEN, catnr. 19901) were added to the lysis buffer, and samples were disrupted in a FastPrep FP24 instrument with the following program: 1 minute heating at 6.5 m/sec, 1-minute incubation at 4 °C, 1-minute heating 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 (GTGYCCAGCMGTCGGTGA/ GGACTACNVGGGTWTCTAAT), using 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].

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 v1.56 [40] and processed following the DADA2 microbiome pipeline using the paired-end R packages DADA2 [41] and phyloseq [42]. Reads were filtered and trimmed using the parameters truncQ=11, truncLen=c<(130,200), and trimLeft=<(30,30) and then denoised. After removing chimeras, amplicon sequence variants (ASVs) table was classified and taxonomy was assigned using the Ribosomal Database Project (RDP) classifier implemented in DADA2 (RDP classifier v2.16/2016; cut off 91.5). The abundance table was then centered, log-transformed, rarefied to even sequencing depth, and decontaminated. For decontamination, we used the prevalence-based contaminant identification method in the R package decontam [43].

**16S statistical analysis.** All the 16S data analyses were performed using R v3.6.0 and the packages vegan (v2.57) [44], phyloseq (v1.34.0) [44], CoDaSeq (v0.99.6) [47], DESeq2 (v1.30.1) [46], Biostrippers (v2.58.0) [45], rstatix (v0.8.0) [50], 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 averaged. 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 groupings. When applicable, pairwise comparisons were performed using post hoc 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 piperacillin-tazobactam, clinical ventilation, cytokine/chemokine, housekeeping genes), 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 log link 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 taxon in the total cohort was used. This value was used to impute the zeros. dbRDA analyses were performed using the capsule function from vegan, first in univariate analyses with 72 metadata variables.
expression analyses are detailed in the original publication18. fastq data, as well as cell type and subtype assignations for all individual cells, were between species- and strain-level diversity.

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 represent 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 1000x and the average of the all 1000 subsamplings was used as the final strain-level diversity value for each sample, as previously described34. 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 represented 1000x and the subsamplings times were used to see 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 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 COMBatGLOuS consortium. Experimental procedures on BAL samples as well as detailed host single-cell gene expression analyses are detailed in the original publication36. 3’ 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. Trimmmomatic37 (v0.38) was used to trim low quality and discarded short reads. Additionally, Prinseq + (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 microbial artifacts (phage phiX174) were mapped with STAR37 (v2.7.1) and discarded. The remaining unmapped reads were mapped against reference microbe genomes in two-step approaches: pseudomolecules reads using mash38 (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 script) were used to pseudoalign reads to microbial genomes, as described previously39. To identify marker genes for these subtypes, using the function feature Neville tools: an “abundances” table containing the number of reads aligned to each gene from the pre-selected set of reference genomes and a pseudomapping 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 (KO). To remove potential artifacts, two additional filters were applied to the taxonomic table: first, if less than 10 different functions (i.e., 10 different KO) 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. bacteriorial reads were assigned their specific barcodes and UMIs as follows: read IDs from the mapped microbial reads were retrieved from the kallisto pseudomapping (.bam) output using SAMTools (v1.9)40. 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. These assignments were 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 Seurat41 object after preprocessing and quality checking the single-cell cohort, as described previously18. 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.test62. 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, 3,186 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 × 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 × 10⁻⁵). 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 × 10⁻⁹) 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 subtype18 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, has 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. Jost Veenhuis at b.scanner@bioinfo.nl). The single-cell RNA-seq data was first described in a separate publication18 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). 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).

Department of Bioinformatics and Functional Genomics, Flanders Institute for Biotechnology (VIB), Leuven, Belgium.

Department of Bioinformatics and Functional Genomics, Flanders Institute for Biotechnology (VIB), Leuven, Belgium.

Department of Bioinformatics and Functional Genomics, Flanders Institute for Biotechnology (VIB), Leuven, Belgium.
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. Lichtiger, J. et al. Omeprazole Young Adult Cohort and Is It 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. Westblad, 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...
Author contributions
V.L.R., A.C.G., J.W., J.R. designed the study. S.I., 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. J.W. and E.W. 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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021

CONTAGIOUS collaborators
Yannick Van Herck14, Alexander Wilmer13, Michael Casaer12, Stephen Rex15, Nathalie Lorent9, Jonas Yserbyt9,11, Dries Testelmans9,11 & Karin Thevissen16

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