Supplementary Materials

Study population

This prospective cohort study was performed at the Prince of Wales Hospital in Hong Kong, China. All patients were diagnosed with COVID-19 by a positive SARS-CoV-2 on reverse transcriptase polymerase chain reaction test in nasopharyngeal swab, nasal swab, deep throat saliva, sputum, or tracheal aspirate, from Jan 2020 to Feb 2021. Patients were excluded if they were unable to be contacted, re-infected by SARS-CoV-2, declined to participate, or died before the follow-up visit. Data including demographics, clinical and laboratory data were extracted from electronic medical records in the Clinical Management System of the Hospital Authority of Hong Kong. Severity of COVID-19 infection was categorised as Table 1.

Table 1 Definition of severity for COVID-19

Severity	**Definition**
Asymptomatic/Mild	No radiographic evidence of pneumonia
Moderate	Pneumonia was present along with fever and respiratory tract symptoms
Severe	Respiratory rate ≥30/min, oxygen saturation ≤93% when breathing ambient air, or PaO2 / FiO2 ≤300 mm Hg (1 mm Hg = 0.133 kPa)
Critical	There was respiratory failure requiring mechanical ventilation, shock, or organ failure requiring intensive care[1].

Post-acute COVID-19 syndrome (PACS) was defined as having at least one persistent symptom which cannot be explained by alternative diagnosis four weeks after recovery from COVID-19. We assessed the presence of 37 most commonly reported symptoms post COVID[2, 3] at an average of 14 months follow-up after viral clearance.

Controls were recruited during the same study period from the community through advertisement and from the Endoscopy Centre at the Prince of Wales Hospital in subjects who had a normal colonoscopy (stools collected before bowel preparation) with the same collection protocol. We selected aged- and gender-matched controls with similar co-morbidities and standard dietary patterns for comparison of gut microbiota composition between subjects with and without COVID-19 infection. Demographics and co-morbidities of controls were listed in Supplementary Table 3. The exclusion criteria for non-COVID-19 controls were (1) the use of antibiotics in the last 6 months; (2) the use of laxatives or anti-diarrheal drugs in the last 3 months; (3) recent dietary changes (e.g., becoming vegetarian/vegan); (4) known complex

infections or sepsis; (5) known history of severe organ failure (including decompensated cirrhosis, malignant disease, kidney failure, epilepsy, active serious infection, acquired immunodeficiency syndrome); (6) bowel surgery in the last 6 months (excluding colonoscopy/procedure related to perianal disease); (7) presence of an ileostomy/stoma; and (8) current pregnancy.

In total, 310 subjects were recruited in this study (Table 2). All samples from patients with COVID-19 and controls were processed and analysed simultaneously.

Table 2 Clinical characteristics of recruited subjects

	COVID-19 (n=155)	Non-COVID-19 Control (n=155)	*p*
Female, n (%)	82 (52.9)	82(52.9)	1.00
Age, years (IQR)	54 (40-65)	53 (41-63)	0.87
BMI	24 (21-27)	24 (22-28)	0.75
Follow-up, months (IQR)	14 (11-18)	NA	
Severity			
Asymptomatic/Mild	91 (58.7)	NA	
Moderate	30 (19.4)	NA	
Severe	23 (14.8)	NA	
Critical	11 (7.1)	NA	

Stool samples

Stool samples were collected at an average of 14 months follow-up after viral clearance. Subjects on the day of follow-up self-sampled at home and couriered to the hospital within 24 hours of collection. All samples were collected in collection tubes containing preservative media (cat. 63700, Norgen Biotek Corp, Ontario Canada) and stored immediately at −80°C until processing. We have previously shown that data of gut microbiota composition generated from stools collected using this preservative medium was comparable with data obtained from samples that were immediately stored at −80°C[4].

Stool DNA extraction and sequencing

The faecal pellet was added to 1 mL of CTAB buffer and vortexed for 30 seconds, then the sample was heated at 95°C for 5 minutes. After that, the samples were vortexed thoroughly with beads at maximum speed for 15 minutes. Then, 40 μL of proteinase K and 20 μL of RNase A was added to the sample and the mixture was incubated at 70°C for 10 minutes. The supernatant was then obtained by centrifuging at 13,000g for 5 minutes and was added into the Maxwell RSC machine for DNA extraction. Extracted DNA was subject to DNA libraries construction, completed through the processes of end repairing, adding A to tails, purification and PCR amplification, using Nextera DNA Flex Library Preparation kit (Illumina, San Diego, CA). Libraries were subsequently sequenced on our in-house sequencer Illumina NextSeq 550 (150 base pairs paired-end) at the Centre for Microbiota Research, The Chinese University of Hong Kong.

Bioinformatics

Raw sequence data were quality filtered using Trimmomatic V.39 to remove adaptor, low-quality sequences (quality score <20), reads shorter than 50 base pairs. Contaminating human reads were filtering using Kneaddata (V.0.7.2 https://bitbucket.org/biobakery/kneaddata/wiki/Home, Reference database: GRCh38 p12) with default parameters. Following this, microbiota composition profiles were inferred from quality-filtered forward reads using MetaPhlAn3 version 3.0.5. GNU parallel[5] was used for parallel analysis jobs to accelerate data processing. Alpha diversity metrics (Shannon diversity, Chao1 richness) were calculated by using the phyloseq package, version 1.26.0. Species whose average abundance and prevalence was less than 0.1% and 5% were filtered out.

Statistical analysis

All analyses were performed in R 4.0.1. Qualitative and quantitative differences between subgroups were analysed using chi-squared or Fisher's exact tests for categorical parameters and Mann-Whitney test for continuous parameters, as appropriate. Principal Coordinates Analysis (PCoA) was used to visualise the clustering of samples based on their species-level compositional profiles. Associations of specific microbial species with patient parameters were identified using the multivariate analysis by linear models (MaAsLin2) statistical frameworks implemented in the Huttenhower Lab Galaxy instance (http://huttenhower.sph.harvard.edu/galaxy/). PCoA and PERMANOVA are implemented in the vegan R package V.2.5–7.

1. Wu J, Liu J, Zhao X, et al. Clinical characteristics of imported cases of coronavirus disease 2019 (COVID-19) in Jiangsu Province: a multicenter descriptive study. Clinical Infectious Diseases 2020;71:706-712.
2. Lambert N, Corps S, El-Azab SA, et al. COVID-19 Survivors' Reports of the Timing, Duration, and Health Impacts of Post-Acute Sequelae of SARS-CoV-2 (PASC) Infection. medRxiv 2021.
3. Lambert NJ, Corps S. COVID-19 “long hauler” symptoms survey report. 2020.
4. Chen Z, Hui PC, Hui M, et al. Impact of preservation method and 16S rRNA hypervariable region on gut microbiota profiling. Msystems 2019;4:e00271-18.
5. Tange O. Gnu Parallel. DOI: https://doi.org/10.5281/zenodo 2018;1146014.