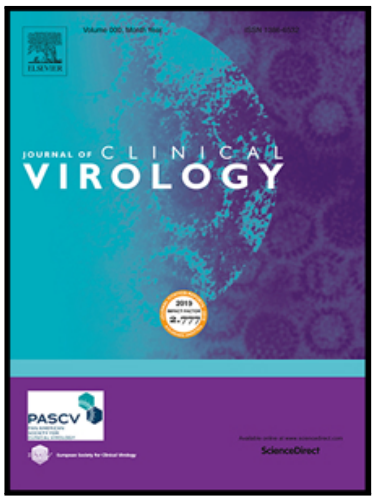

Enrichment of SARS-CoV-2 sequence from nasopharyngeal swabs whilst identifying the nasal microbiome

Abdulrahman Alrezaihi , Rebekah Penrice-Randal , Xiaofeng Dong , Tessa Prince , Nadine Randle , Malcolm G. Semple , Peter J.M. Openshaw , Tracy MacGill , Todd Myers , Robert Orr , Samo Zakotnik , Alen Suljič , Tatjana Avšič-Županc , Miroslav Petrovec , Miša Korva , Waleed AlJabr , Julian A. Hiscox , ISARIC4C Investigators

PII: S1386-6532(23)00243-3
DOI: https://doi.org/10.1016/j.jcv.2023.105620
Reference: JCV 105620

To appear in: *Journal of Clinical Virology*

Received date: 8 June 2023
Revised date: 6 November 2023
Accepted date: 18 November 2023

Please cite this article as: Abdulrahman Alrezaihi , Rebekah Penrice-Randal , Xiaofeng Dong , Tessa Prince , Nadine Randle , Malcolm G. Semple , Peter J.M. Openshaw , Tracy MacGill , Todd Myers , Robert Orr , Samo Zakotnik , Alen Suljič , Tatjana Avšič-Županc , Miroslav Petrovec , Miša Korva , Waleed AlJabr , Julian A. Hiscox , ISARIC4C Investigators, Enrichment of SARS-CoV-2 sequence from nasopharyngeal swabs whilst identifying the nasal microbiome, *Journal of Clinical Virology* (2023), doi: https://doi.org/10.1016/j.jcv.2023.105620

Highlights

- Metagenomic analysis of clinical samples from patients with COVID-19
- Identification of SARS-CoV-2 lineages
- Identification of seasonal coronaviruses
- Identification of the nasal microbiome
- Sequence-independent, single-primer amplification (SISPA)
- Oxford nanopore long read length sequencing of coronaviruses

Enrichment of SARS-CoV-2 sequence from nasopharyngeal swabs whilst identifying the nasal microbiome

Abdulrahman Alrezaihi[1,2], Rebekah Penrice-Randal[1], Xiaofeng Dong[1], Tessa Prince[1], Nadine Randle[1], Malcolm G. Semple[1,3,4], Peter J. M. Openshaw[5], ISARIC4C Investigators, Tracy MacGill[6], Todd Myers[6], Robert Orr[6], Samo Zakotnik[7], Alen Suljič[7], Tatjana Avšič-Županc[7], Miroslav Petrovec[7], Miša Korva[7], Waleed AlJabr[1,8], Julian A. Hiscox[1,3,9].

[1]University of Liverpool, Liverpool, UK.

[2]King Saud University, Riyadh, Saudi Arabia.

[3]NIHR Health Protection Research Unit in Emerging and Zoonotic Infections, Liverpool, UK.

[4]Alder Hey Children's Hospital, Liverpool, UK.

[5]Imperial College London, London, UK.

[6]Office of Counterterrorism and Emerging Threats, U.S. Food and Drug Administration, Silver Spring, USA.

[7]University of Ljubljana, Ljubljana, Slovenia.

[8]King Fahad Medical City, Riyadh, Saudi Arabia.

[9]Agency for Science, Technology and Research (A*STAR), Singapore.

Abstract

Simultaneously characterising the genomic information of coronaviruses and the underlying nasal microbiome from a single clinical sample would help characterise infection and disease. Metatranscriptomic approaches can be used to sequence SARS-CoV-2 (and other coronaviruses) and identify mRNAs associated with active transcription in the nasal microbiome. However, given the large sequence background, unenriched metatranscriptomic approaches often do not sequence SARS-CoV-2 to sufficient read and coverage depth to obtain a consensus genome, especially with moderate and low viral loads from clinical samples. In this study, various enrichment methods were assessed to detect SARS-CoV-2, identify lineages and define the nasal microbiome. The methods were underpinned by Oxford Nanopore long-read sequencing and variations of sequence independent single primer amplification (SISPA). The utility of the method(s) was also validated on samples from patients infected seasonal coronaviruses. The feasibility of profiling the nasal microbiome using these enrichment methods was explored. The findings shed light on the performance of different enrichment strategies and their applicability in characterising the composition of the nasal microbiome.

Introduction

Respiratory disease is often multifactorial and can be caused by different pathogens interacting synergistically and disrupting microbial ecology [1, 2]. The microbiome sampled from a nasopharyngeal swab may be reflective of the respiratory microbiome and can be obtained from different sampling strategies including sequencing, culture and arrays [3]. During the first wave of the COVID-19 pandemic, respiratory viruses other than SARS-CoV-2 were often excluded in diagnostics due to prioritising identification of COVID-19 in patients [3]. During a pandemic caused by coronaviruses or analysing legacy samples or in 'peacetime', being able to characterise the totality of coronavirus infection from a single sample would be advantageous. This would be to simultaneously derive coronavirus genome sequence and investigate how a nasal microbiome/co-infection may influence disease and outcome, especially if sample sets or amounts are limited.

SARS-CoV-2 has a positive sense single stranded RNA genome of approximately 29,900 nucleotides and replicates in the cytoplasm of an infected cell. The 5' two thirds of the genome is immediately translated into two polyproteins that are proteolytically cleaved to generate a variety of proteins including those involved in replication [4, 5]. The remaining one third of the genome is expressed through the transcription of a nested set of subgenomic RNAs (sgmRNAs), that share a 5' leader sequence with the genome and polyA tail [4]. Control of sgmRNA transcription is in part due to the transcription regulatory sequence (TRS) that precedes each gene along the genome [6]. The general architecture of a sgmRNA is 5' to 3', the leader sequence followed by the TRS (called a leader-TRS), followed by the gene to be translated and then other genes (depending on the sgmRNA), a non-coding reading and a polyA tail. The leader sequence is also found at the 5' end of the

genomic RNA. Detection of SARS-CoV-2 and viral load information from a clinical specimen can involve identification and quantification of the genome/sgmRNAs [7]. The leader-TRS gene junction is also a unique sequence feature that can be used to identify and quantify subgenomic and genomic RNA, particularly using sequencing information (e.g. [8-10]). The leader-TRS nucleoprotein gene junction is normally the most abundant because during active infection of a cell by a coronavirus the sgmRNA encoding the nucleoprotein is the most abundant [4, 6].

In healthy individuals microbial communities exist in the upper and lower respiratory tract and can consist of the phylum Bacteroidetes, Firmicutes, Proteobacteria, and Actinobacteria (reviewed in [11]). Information on this can be inferred by elucidating the nasal microbiome. Disruption of the respiratory microbiome can be associated with disease such as translocation of gut bacteria to the lung and association with acute respiratory distress syndrome (ARDS) [12]. This imbalance or disruption of the respiratory microbiome (or any microbiome) and association with disease is called dysbiosis. The respiratory microbiome may be perturbed during coronavirus infection and other co-infections requiring clinical management can be present [13]. During the first wave of the COVID-19 pandemic (at least in the UK experience), respiratory viruses other than SARS-CoV-2 were often excluded in diagnostics due to prioritising identification of COVID-19 in patients [3]. Several studies have characterised the nasal microbiome in patients with SARS-CoV-2 with inconsistent results (e.g. [14-16]). Although, reduced abundance of Corynebacterium has been associated with anosmia in patients with COVID-19 [17] and a pattern of dysbiosis has been reported (e.g. [18-21]).

A broad range of pathogens (viruses, bacteria, fungi, and parasites) can be identified within a clinical sample using random amplification and shotgun sequencing [13, 22-25]. Detection of RNA transcripts through metatranscriptomics, also gives an indication of the biological activity of the pathogen that is present [22]. One of the most prominent approaches for random amplification is sequence-independent single primer amplification (SISPA). This approach can be effective as an investigative tool for identifying multiple infectious diseases [23, 24] and elucidating complex microbiomes [22]. The analysis of legacy samples from the COVID-19 pandemic continues to provide new insights into the evolution and spread of SARS-CoV-2 as well as the disease profile of different variants. Maximising information from single sample would be advantageous to characterise infection. In this study, a metatranscriptomic approach based around SISPA was optimised with Oxford Nanopore sequencing to provide detailed sequence/lineage information on SARS-CoV-2 as well as provide data on the underlying active nasal microbiome and validated for use in other human coronavirus infections.

Methods

RNA Extraction and Preparation

Nasopharyngeal swabs were collected into a viral transport medium from patients diagnosed with SARS-CoV-2 (n=12) and other seasonal human coronaviruses: HCoV-229E (n=2), HcoV-HKU1(n=1) and HcoV-OC43 (n=2) (identification through RT-PCR in all cases). RNA was isolated using either a QIAamp Viral RNA Mini Kit (Qiagen, Manchester, UK) or Trizol LS (Invitrogen, UK). As an internal control for identifying viral RNA (genome and sgmRNAs) for comparing sequencing methodologies total RNA was purified from SARS-CoV-2 infected Vero cells using the Qiagen RNA Mini Kit following AVL inactivation. RNA samples were treated with Turbo Dnase (Invitrogen). SARS-CoV-2 (as with other many other viruses) can be grown to high titre in Vero cells [26] and this provides enough viral RNA for sequencing.

Sequencing

Four methodologies were used. First, was standard standard SISPA (referred to as SISPA), the second included a primer to amplify viral targets containing the SARS-CoV-2 leader sequence (SISPA-L). This would include enrichment of the 5’ region of the genome and sgmRNAs. The third and fourth methodologies included the leader sequence but were also based on two different tiled amplicon methodologies to sequence SARS-CoV-2, ARTIC [27] (SISPA-ARTIC-L) or rapid sequencing long amplicon RSLA (SISPA-RSLA-L) [28]. ARTIC sequencing was based on generating amplicons to cover the SARS-CoV-2 with average length of 400 by and RSLA to cover the SARS-CoV-2 (or MERS-CoV) genome with an average length between 1000 bp and 3000 bp [13, 28]. RNA was reverse-transcribed with SuperScript IV Reverse Transcriptase (Invitrogen) using Sol-PrimerA (5’-

GTTTCCCACTGGAGGATA-N9-3'). Following cDNA synthesis, PCR amplification was performed, and PCR products were subsequently purified at a 1:1 ratio with AMPure XP beads (Beckman Coulter, High Wycombe, UK). The cycling conditions were as follows; 98°C for 30 s, followed by 30 cycles of 98°C for 10 s, 54°C for 30 s, and 72°C for 60s, with a final extension at 72°C for 10 min. The library was prepared as per the sequencing by ligation protocol SQK-LSK109 with Native Barcoding Expansion 1-12 (EXP-NBD104) and 13-24 (EXP-NBD114) for multiplexing (Oxford Nanopore, Oxford, UK).

Bioinformatics

Fast5 files generated by the GridION were base called using Guppy basecaller filtering low quality reads (Oxford Nanopore, Oxford, UK). For assembly of the SARS-CoV-2 genomes, Minimap2 (v. 2.17-r941) was used to align fastq sequences to the SARS-CoV-2 isolate Wuhan-Hu-1 reference genome (NC_045512.2), and to human coronavirus (NC_002645.1 229E, NC_006577.2 HKU1, NC_005147.1 OC43) using the -ax map-ont parameters. Samtools (v.1.10) was used to sort and index alignment files, and Picard (v.2.23.4) was used to mark duplicates. Mapped reads were compared between different conditions using the non-parametric Wilcoxon Rank-Sum Test (Mann-Whitney U Test) due to non-normal distribution (performed in R). The LeTRS tool was used to identify coronavirus gmRNAs (https://github.com/Hiscox-lab/LeTRS) [8]. Fastq files were then trimmed using NanoFilt [29], before classifying reads to viral, bacterial, and fungal species using kraken2 [30]. The local database used in this study for metatranscriptomics was constructed by downloading and integrating various taxonomic libraries from (https://github.com/DerrickWood/kraken2). The databases downloaded included bacterial, viral, fungal, human, archaeal, and protozoal reference sequences. The kraken2 version used in this study was 2.0.9-beta. Each library

was downloaded using the kraken2-build --download-library command with appropriate parameters. The resulting database was then constructed using the kraken2-build --build command. Kraken-biom was utilised to convert kraken reports into biom format which was imported into Rstudio using the Phyloseq package. SAMtools and BCFtools [31] were utilized to extract SARS-CoV-2 mapped reads and consensus fasta files were generated with the Seqtk tool. We noted that the Kraken analysis can return many organism identifications. In order to down select to those most relevant and likely to be positive several stages were used. Reads mapping to fungi and yeast were not included in the final analysis due to insufficient coverage and a distinct minimizer cut off (confidence level) with kraken that was used (https://github.com/DerrickWood/kraken2/wiki/Manual#distinct-minimizer-count-information). For interest, reads that mapped either to fungi or yeast can be derived from the sequencing that was uploaded as part of this project (see below).

The nasal microbiome was evaluated or described using a number of different parameters. The alpha diversity (number of species) of the nasal microbiome was evaluated by deriving the Observed and Chao1 indices, which focus on the richness level. The Observed index calculated the number of unique microbial species observed in a sample, while the Chao1 index estimated the total number of microbial species present in the sample, regardless of their abundance. The Shannon index, on the other hand, took into account both the richness and evenness of microbial species present in the sample, providing a more comprehensive measure of diversity. For the alpha diversity analysis, different statistical tests were applied using R and the 'phyloseq' package. The comparisons were performed using the pairwise t.test function for the Observed richness and Chao1 indices, while the pairwise.wilcox.test function in R was used for the Shannon index. The level of significance:

one asterisk (*) represents $p < 0.05$, two asterisks (**) represent $p < 0.01$, and three asterisks (***) represent $p < 0.001$. Note in this study with default parameters in the Observed index, approximately 2000 species were identified in a clinical sample, whereas when the cut-off (0.05) was applied less than 500 species were identified. The sequencing data in this study has been deposited with BioProject ID (PRJNA1000473) (https://www.ncbi.nlm.nih.gov/bioproject/1000473) where the user can use this to input into Kraken for their own analysis.

Results

SISPA was used and modified to obtain information about the RNA species within a clinical sample. This facilitated the study of the composition of the nasal microbiome and viral genome sequences. During analysis, there were several interrelated factors to consider, including mapped reads, genome coverage and read depth.

Enrichment of SISPA based methodology to enhance viral genome sequence using samples from patients with COVID-19

To enhance coronavirus genomic data in the context of a clinical sample, the SISPA approach was used as basis with three specific modifications. The ability of SISPA, SISPA-L, SISPA-ARTIC-L and SISPA-RSLA-L as methodologies to sequence SARS-CoV-2 were compared by using RNA purified from cells infected in culture with SARS-CoV-2 generating high viral loads (Fig. 1A). The data indicated that SARS-CoV-2 sequence was derived from the four different methodologies and that SISPA-ARTIC-L generally had the greatest number of mapped reads (Fig. 1A). Next these methodologies were compared on nasopharyngeal samples taken from patients with COVID-19 with a range of viral loads (Ct 22 to 36) (Fig. 1A). There was a greater number of mapped reads for SARS-CoV-2 in clinical samples with higher viral loads, and these were generally associated with SISPA-ARTIC-L (Fig. 1A).

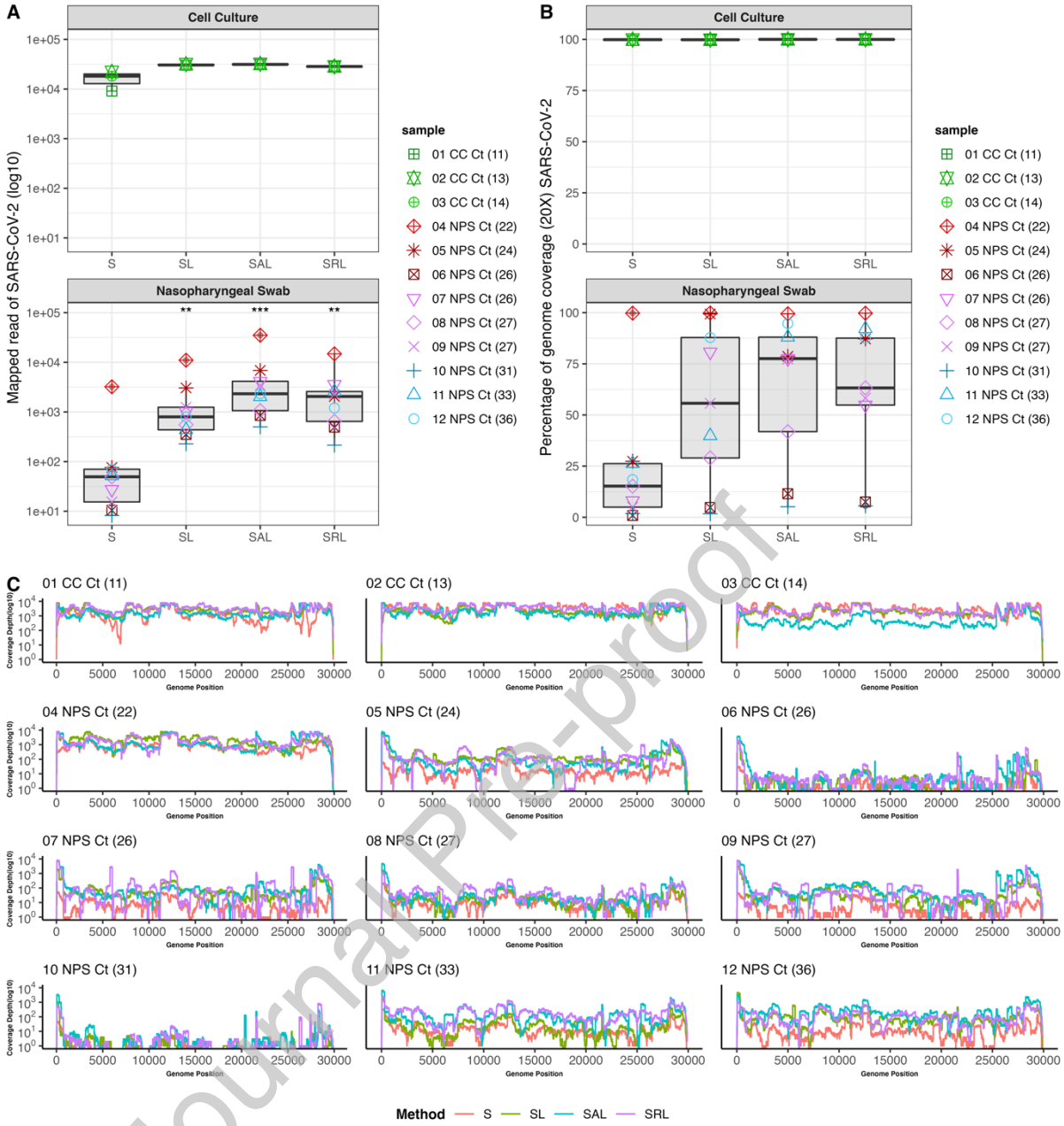

Fig. 1. (A). Comparison of mapped reads on the SARS-CoV-2 genome using different amplification methodologies sequenced by long read length Oxford Nanopore. The methodologies were SISPA (S), SISPA-L (SL), SISPA-ACTIC-L (SAL) and SISPA-RSLA-L (SRL) (x-axis). Mapped reads using the four methodologies were compared between RNA extracted from cells infected in culture (CC notation) and RNA extracted from nasopharyngeal swabs from patients with COVID-19 (NPS notation). The y-axis shows the number of mapped reads, with the horizontal line within each boxplot denotes the median value, and the box

represents the interquartile range. The key indicates the sample type and the figure in parenthesis is the viral load. (B) The percentage of SARS-CoV-2 genome coverage (20x) in the different sample types sequenced by four different methodologies. Percentage genome coverage using the four methodologies were compared between RNA extracted from cells infected in culture (CC notation) and RNA extracted from nasopharyngeal swabs from patients with COVID-19 (NPS notation). The viral load/Ct value for each sample are given in the key to each sample in parenthesis. (C) The genome coverage depth of SARS-CoV-2 at each nucleotide position (x-axis). Shown is sequencing data from either RNA extracted from cells infected in culture (CC notation) and RNA extracted from nasopharyngeal swabs from patients with COVID-19 (NPS notation). The four different methodologies, SISPA (S), SISPA-L (SL), SISPA-ACTIC-L (SAL) and SISPA-RSLA-L (SRL) are colour coded and coverage depth is shown on the y-axis. The viral load/Ct value for each sample is denoted in parenthesis.

Next, the percentage of mapped reads to the SARS-CoV-2 genome was compared between the different sequencing methodologies (Fig. 1B). A cut-off value of a sequence read depth of 20 was applied. For low Ct values there was 100% genome coverage (Fig. 1B). For the high Ct values there was low % coverage of the genome. In general, SISPA-ARTIC-L gave higher genome coverage compared to the other methodologies.

Comparison of depth of coverage and genome position between the samples indicated that with higher viral loads there was complete genome coverage and high read depth – irrespective of the methodology used. With sequence from cell culture and one clinical sample (Ct 22) there was complete coverage of the SARS-CoV-2 genome and equivalent sequence read depth (Fig. 1C). For samples with lower viral loads, sequence read depth was

not evenly distributed across the genome. When comparing the four methodologies, SISPA-RSLA-L, gave slightly more read depth and coverage in samples with lower viral load.

Characterisation and comparison of SARS-CoV-2 gene expression patterns

SARS-CoV-2 transcription patterns can be distinguished from sequence data. The methodologies were evaluated in their ability to derive gene expression profiles using LeTRS [8]. All the major known leader-TRS gene junctions were identified (Fig. 2), with the leader-TRS nucleoprotein gene junction usually being the most abundant. In general, when comparing the sample count of known leader-TRS gene junctions between conditions, the SISPA-ARTIC-L methodology provided the highest count (Fig. 2). In some cases, sgmRNAs were not detected from the sequencing data and this included the ORF10 sgmRNA (Fig. 2). ORF10 is a unique feature of SARS-CoV-2. This may reflect the lower abundance of this transcript in cells infected in culture [4].

Fig.2. Abundance of ten known leader-TRS junctions of SARS-CoV-2 (Leader-S to Leader-ORF10) and four potential novel leader-TRS junctions identified from the sequencing data (right hand side) using LeTRSs – a bioinformatic based tool developed to detect and quantify defined leader-TRS gene junctions of SARS-CoV-2 (or other coronaviruses from sequencing data) [8]. This provides a measurement for the abundance of each sgmRNA made in a cell/present in a clinical sample using sequence data, rather than more conventional

techniques such as northern blot or metabolic labelling. The four different methodologies, SISPA (S), SISPA-L (SL), SISPA-ACTIC-L (SAL) and SISPA-RSLA-L (SRL) are colour coded and presented on the x-axis. The y-axis is the normalised count for each leader-TRS junction in the cell culture (CC notation) and clinical samples (NPS notation). Each leader-TRS gene junction is presented in the order of the appropriate gene along the genome. ORF10 is the 3' most gene so far identified in SARS-CoV-2. In an active infection the nucleoprotein leader-TRS gene junction would be the most abundant.

Defining lineage information on SARS-CoV-2

Lineage information on SARS-CoV-2 can be used to track spread and evolutionary change in the virus. Using the Phylogenetic Assignment of Named Global Outbreak Lineages (Pangolin) algorithm (v3.1.16) [32], sequencing data from the four methodologies were evaluated in the ability to assign a PANGO lineage to SARS-CoV-2. Consistent lineages were assigned for all four methodologies in samples with higher viral loads (Table 1). However, in samples with lower viral loads (e.g, 10 NPS) PANGO lineages were not obtained or varied in degree of assignment (Table 1). In general, the enriched SISPA methodologies provided more detailed lineage information.

Comparison of the different sequencing methodologies for enrichment of SARS-CoV-2 to provide context for the background nasal microbiome

The rationale for comparing four methodologies for sequencing SARS-CoV-2 from nasopharyngeal swabs was also to obtain information on the nasal microbiome. This took in various factors to distinguish the nasal microbiome including relative abundance and alpha

diversity (richness and evenness). To identify and quantify the composition of the nasal microbiome, sequencing reads were assigned a taxonomy with kraken2 [33] (Fig. 3).

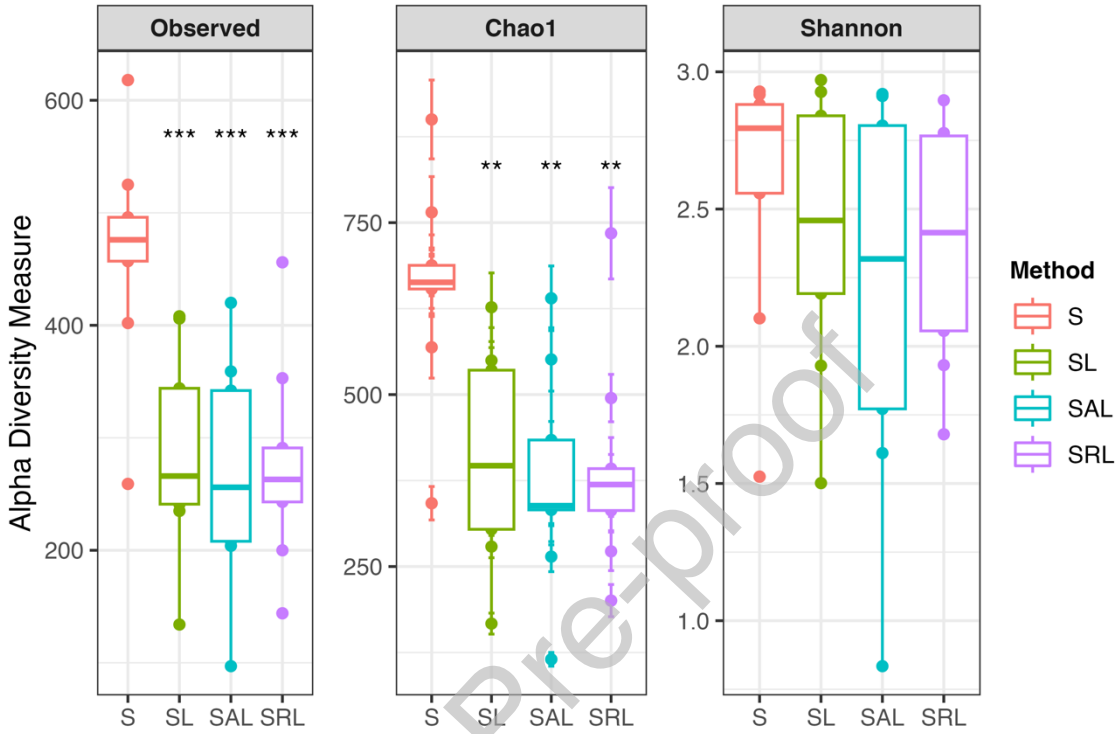

Fig. 3. Comparison of the Alpha diversity between the SISPA (S), SISPA-L (SL), SISPA-ACTIC-L (SAL) and SISPA-RSLA-L (SRL) methodologies (colour coded in the key) using the observed, Chao1 and Shannon index as parameters. The four methodologies are shown as the x-axis and the Alpha diversity measure is shown in the y-axis. The horizontal line denotes the median value.

The highest abundance of SARS-CoV-2 across the samples was identified with the SISPA-ARTIC-L methodology (Fig. 4A). This was in line with the reference-based mapping approach. The richness of the microbiome in the samples was compared using the Observed and Chao1 indexes to evaluate the number of distinct species present. For these parameters the results showed significant differences in richness between SISPA and the other three

enrichment methods as determined by Pairwise t tests. However, there was no significant difference in species evenness and richness between the methodologies when comparing the Shannon index (Fig. 3).

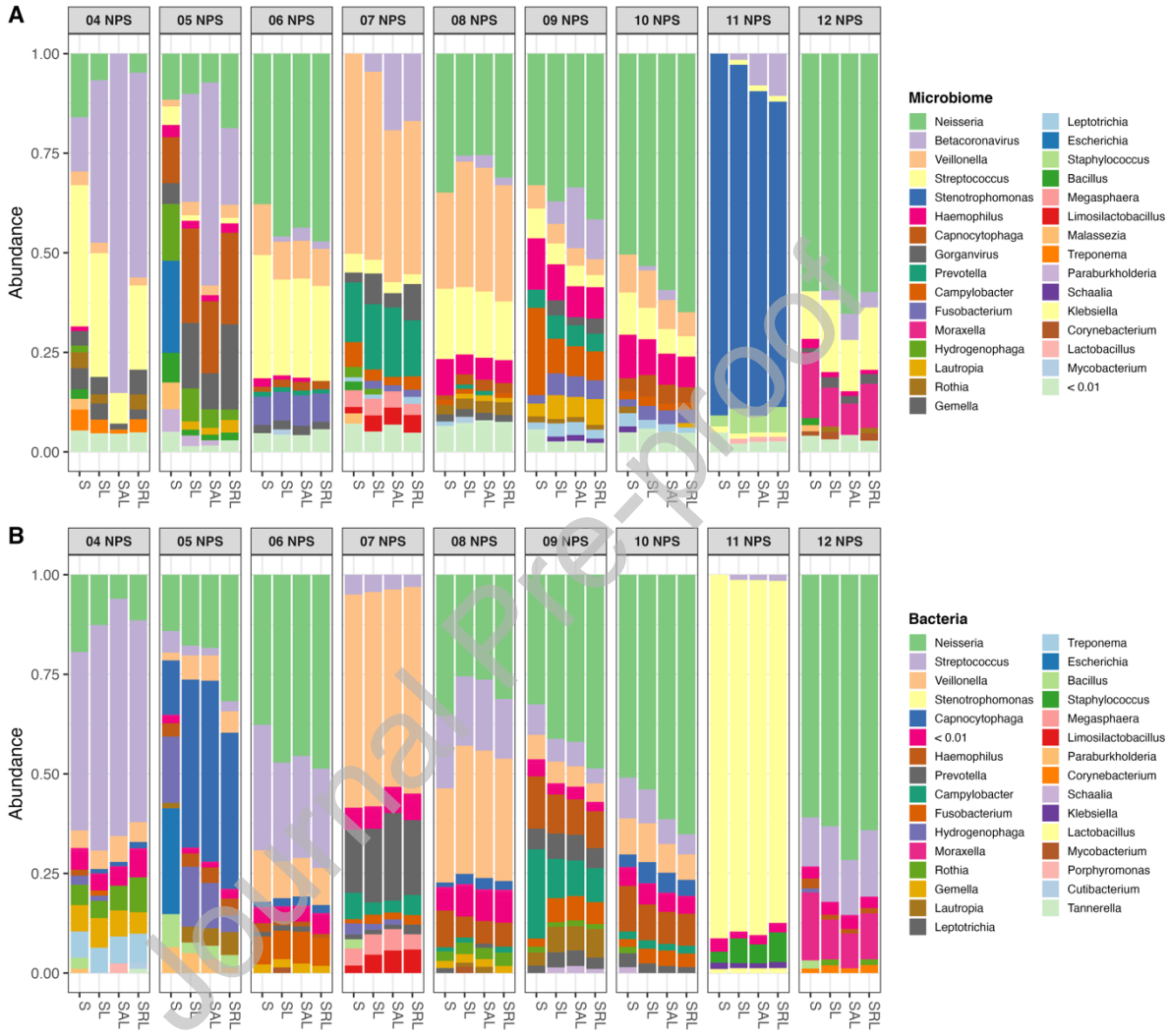

Fig.4. Bar plots of the top taxa (indicated and colour coded to the right) plotted by genus identified using the four different methodologies; SISPA (S), SISPA-L (SL), SISPA-ACTIC-L (SAL) and SISPA-RSLA-L (SRL) shown on the x-axis together with the ID of the clinical sample. The y-axis is the relative abundance of each genus. Shown in each panel: (A) the total microbe encompassing the top 50 taxa, (B) the top 50 bacterial taxa only.

The viral genome enrichment methodology appeared to impact the abundance of the nasal microbiome, especially in samples with higher viral loads (Fig. 4A). To investigate this, the nasal microbiome was sub-divided into the relative abundance of bacterial (Fig. 4B). As an example, sample 04 NSP was considered by excluding viral reads from the analysis. In this case the Chao1 index indicated that the SISPA methodology provided the highest measurement for alpha diversity compared to the other viral genome enrichment strategies (Sup. 1A). However, in this situation, the Shannon index indicated that the alpha diversity was comparable for each method (Sup. 1B).

The data indicated that SISPA alone was optimal for capturing information on the nasal microbiome. However, the SISPA-L methodology was optimal for recovering sequence information for a specific known virus (SARS-CoV-2) whilst still maintaining the ability to capture the diversity and scope of the underlying nasal microbiome in a clinical sample.

Enrichment of SISPA for seasonal coronaviruses

The enhancement of nasal microbiome and specific virus sequencing has utility not just for SARS-CoV-2 in clinical samples. Particularly, where a coronavirus may have been identified through an array-based approach, but complete genomic information is not yet available for accurate amplicon-based sequencing. Therefore, to evaluate the ability to capture information on samples with a known (coronavirus) and unknown (microbiome) the SISPA and SISPA-L methodologies were compared in samples with a known seasonal human virus but unknown microbiome. In this case, five nasopharyngeal samples that tested positive for human seasonal coronaviruses (human coronavirus 229E – HCoV-229E), human coronavirus HKU1 – HCoV-HKU1 and human coronavirus OC43 – HCoV-OC43) were used. Using

referenced based mapping approaches for the seasonal coronavirus, in general, the SISPA-L methodology provided a greater proportion of reads mapping to the appropriate coronavirus compared to SISPA alone (Fig. 5A). When comparing coverage across the viral genome (Fig. 5B), again the SISPA-L methodology was more efficient at providing coverage. Interestingly, in the sample containing HCoV-HKU1, there was a greater proportion of reads that mapped to the 3' end of the genome, potentially indicative of an active infection – where the sgmRNAs were expressed. Non-referenced based mapping approaches using kraken2 provided information on potential viruses and bacteria that were present in the nasal microbiome (Fig. 6A). The SISPA-L methodology was effective in identifying the highest abundance of HCoV-229E (alpha coronavirus) and HCoV-HKU1, HCoV-OC43 (beta coronavirus) across the samples, which was consistent with the reference-based mapping approach. To assess the richness of the nasal microbiome in the samples, the Observed and Chao1 indexes were used. The results revealed significant differences in richness between SISPA and SISPA-L ($p < 0.05$), as determined by a non-parametric Wilcoxon rank-sum test. However, no significant difference was observed in species evenness and richness between the methodologies when comparing the Shannon index (Fig. 6B).

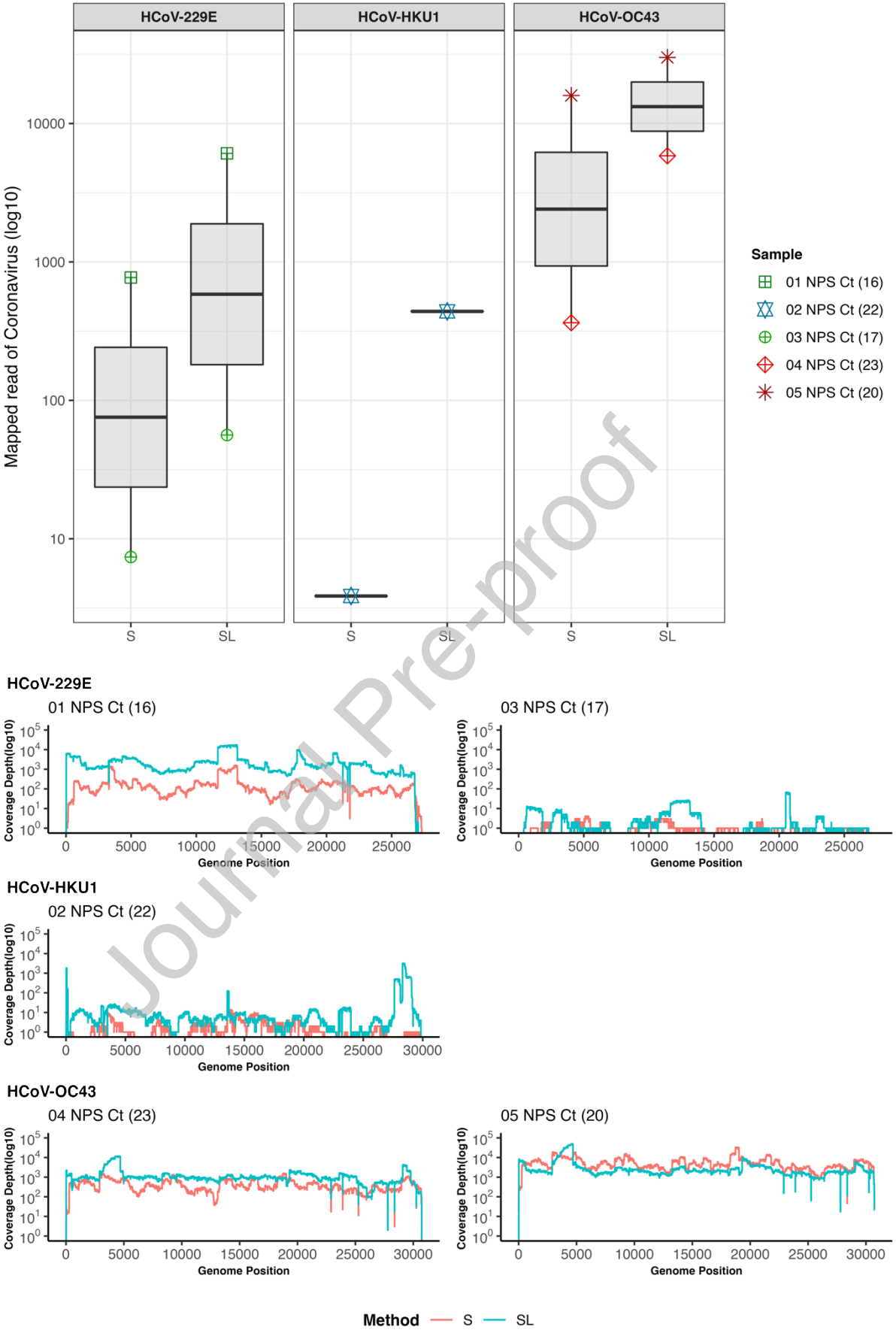

Fig. 5. (A) and (B) Analysis of mapped reads and genome coverage depth for seasonal human coronaviruses showcasing the outcomes obtained from sequencing data generated through both SISPA and SISPA-L approaches with virus-specific leader primers. (A) Comparison of mapped reads on the genomes of seasonal human coronaviruses (HCoV-229E, HCoV-HKU1 and HCoV-OC43) from sequencing data generated either through SISPA or a SISPA-L (x-axis) using a specific leader primer for each virus. The y-axis shows the number of mapped reads, with the horizontal line within each boxplot denotes the median value, and the box represents the interquartile range. Sequence data from five clinical samples (indicated to the right) was compared. (B) Comparison of the genome coverage depth for three seasonal human coronaviruses at each position along the genome. Shown is the sequencing data from nasopharyngeal swabs (NPS) (indicated for each figure). Genome position is shown on the x-axis and read/coverage depth on the y-axis. Coverage is color coded for each method.

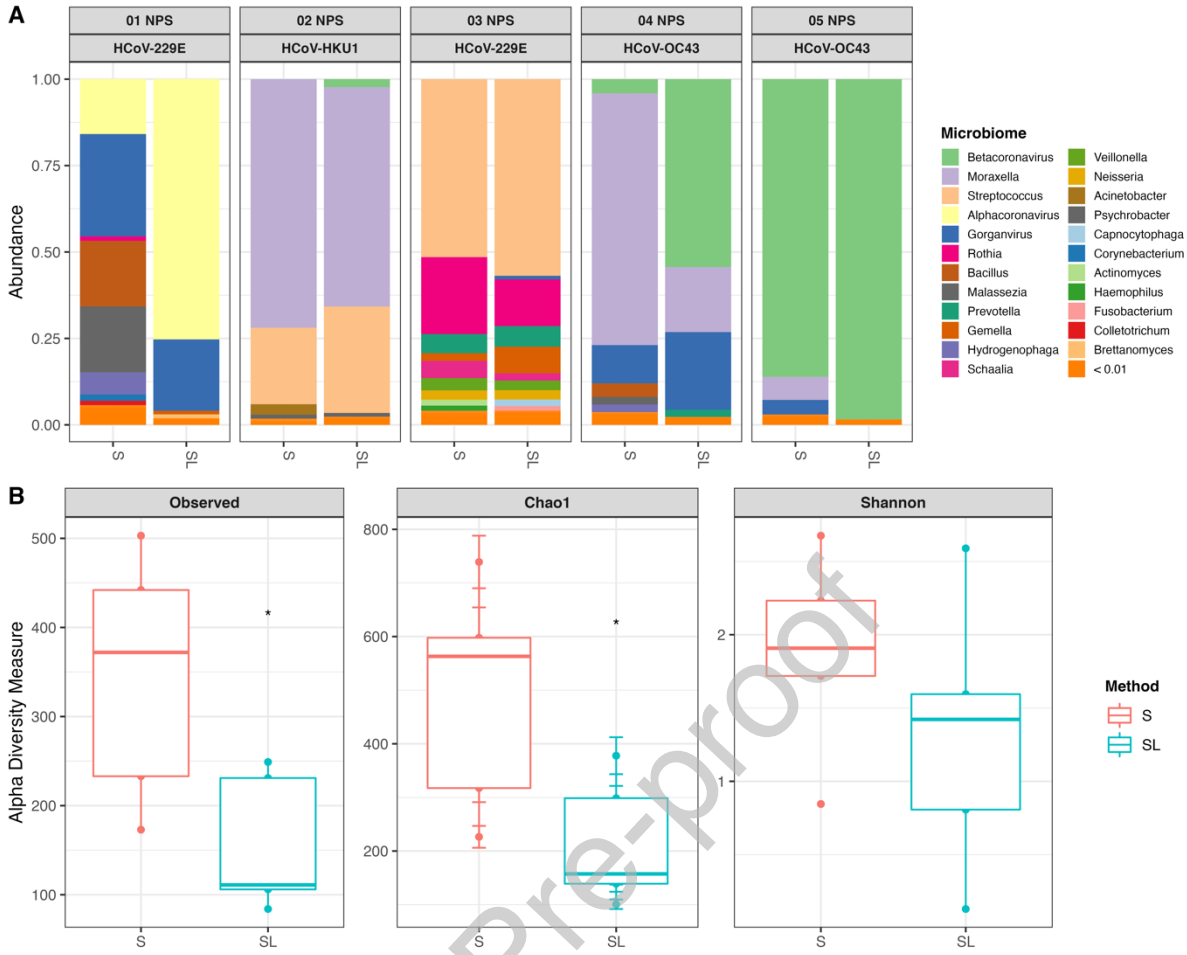

Fig. 6. (A) Bar plots of the total nasal microbiome encompassing the top 50 taxa (indicated and colour coded to the right) plotted by genus identified using the two different methodologies; SISPA (S), SISPA-L (SL) shown on the x-axis together with the ID of the clinical sample. The y-axis is the relative abundance of each genus. (B) Comparison of the Alpha diversity between the SISPA (S) and SISPA-L (SL) methodologies (colour coded in the key) using the observed, Chao1 and Shannon index as parameters. The two methodologies are shown as the x-axis and the Alpha diversity measure is shown in the y-axis.

Discussion

In this study, different enrichment methods for targeting and analysing coronavirus genomes in clinical samples and providing information on the nasal microbiome were evaluated. Many clinical samples have low to moderate viral load and low-quality RNA, which makes it challenging to obtain sequence information. We postulated that an enhanced metatranscriptomic approach could be used to provide both coronavirus genomic information and the active microbiome. Coronaviruses have a unique sequence feature (the leader) that provides an opportunity to increase genomic coverage in the context of a metatranscriptomic strategy by spiking in an appropriate primer to amplify target viral sequence.

The three enrichment methodologies, SISPA-L, SISPA-ARTIC-L, and SISPA-RSLA-L provided a significant increase in the level of mapped reads and coverage for SARS-CoV-2 when compared with SISPA only. However, enhancing viral specific reads resulted in a decrease in the number of reads that could be used to characterise the background microbiome. SISPA-ARTIC-L demonstrated the highest enrichment for viral reads, while SISPA detected the highest number of species in the microbiome. A compromise was the SISPA-L methodology that could provide sufficient viral read depth to identify a PANGO lineage for SARS-CoV-2 whilst covering the microbiome. To evaluate this further, SISPA and SISPA-L was used to characterise seasonal human coronaviruses and the nasal microbiome from clinical samples. Again, enhanced viral read depth was obtained with SISPA-L whilst still providing detail on the nasal microbiome.

In general, no matter what the metatranscriptomic approach, the clinical samples from patients with COVID-19 had higher abundances of *Neisseria* (Fig. 4). Although in some patients this was *Escherichia* or *Stenotrophomonas*. Dysbiosis [20] and *Neisseria* signatures in the oropharyngeal microbiome have been associated with severe COVID-19 [34]. In the wider perspective during containment policies (e.g. movement restrictions/'lock downs' the general incidence of disease caused by bacteria such as Neisseria meningitidis, Streptococcus pneumoniae and Haemophilus influenzae decreased [35]. This may in turn may have affected different nasal microbiomes identified in patients with COVID-19 – as well as when in the history of the patient the sample was taken – particularly in the presence of antibiotics [36].

No doubt as legacy samples from the COVID-19 pandemic continue to be analysed and in ongoing cases of patients with SARS-CoV-2, the ability to define detailed sequence information about the virus and also the context of the background nasal/respiratory microbiome will be important. Particularly in considering future antibiotic stewardship and whether a particular component(s) of this microbiome requires clinical management. The ability to identify coronaviruses in routine samples and a future coronavirus as a Disease X, will help characterise pathogenicity.

Acknowledgements

We would like to thank members of the Hiscox Laboratory and the ISARIC4C team, Lance Turtle (University of Liverpool) and J. Kenneth Baillie (University of Edinburgh) for supporting this work.

Funding

This work was funded by U.S. Food and Drug Administration Medical Countermeasures Initiative contract (75F40120C00085) awarded to JAH. The article reflects the views of the authors and does not represent the views or policies of the FDA. This work was also supported by the MRC (MR/W005611/1) G2P-UK: A national virology consortium to address phenotypic consequences of SARS-CoV-2 genomic variation (co-I JAH). AA is funded by a studentship from King Saud University, Saudi Arabia. The ISARIC4C sample collection and sequencing in this study was supported by grants from the Medical Research Council (grant MC_PC_19059), the National Institute for Health Research (NIHR; award CO-CIN-01) and the Medical Research Council (MRC; grant MC_PC_19059). JAH, MGS and LT are supported by the NIHR Health Protection Research Unit (HPRU) in Emerging and Zoonotic Infections at University of Liverpool in partnership with the UK Health Security Agency (UK-HSA), in collaboration with Liverpool School of Tropical Medicine and the University of Oxford (award 200907). LT is supported by a Wellcome Trust fellowship [205228/Z/16/Z].

We declare that we have no conflict of interest.

Ethics Statement

For clinical samples gathered under the auspices of ISARIC4C patients were recruited under the International Severe Acute Respiratory and emerging Infection Consortium (ISARIC)

Clinical Characterisation Protocol CCP (https://isaric4c.net/) by giving informed consent. Ethical approval was given by the South Central - Oxford C Research Ethics Committee in England (Ref 13/SC/0149), the Scotland A Research Ethics Committee (Ref 20/SS/0028), and the WHO Ethics Review Committee (RPC571 and RPC572, 25 April 2013). The use of clinical respiratory samples for seasonal coronaviruses was approved by the National Medical Ethics Committee of Slovenia (No. 0120-211/2020/7 and 0120-56/2023/3).

Author contribution statement

Conceptualization: AA, TMac, TM, RO, WA and JAH. Resources: MGS, PJMO, JKB, ISARIC4C Investigators, LT, SZ, AS, TA-Z, MP and MK. Data curation: AA, RP-R and XD. Software: AA, RP-R and XD. Formal analysis: AA. Supervision: WA, TA-Z, MP, MK and JAH. Funding acquisition: WA, MGS, PJMO, JKB, LT and JAH. Validation: AA. Investigation: AA, RP-R, XD, TP, NR, SZ and AS. Visualization: AA. Methodology: AA and JAH. Writing – original draft: AA and JAH. Writing – review and editing – AA, SZ, RP-R, MGS and JAH. Project administration: TMac, TM, RO and JAH.

References

[1] Budden KF, Shukla SD, Rehman SF, Bowerman KL, Keely S, Hugenholtz P, et al. Functional effects of the microbiota in chronic respiratory disease. Lancet Respir Med. 2019;7:907-20.

[2] Natalini JG, Singh S, Segal LN. The dynamic lung microbiome in health and disease. Nat Rev Microbiol. 2023;21:222-35.

[3] Russell CD, Fairfield CJ, Drake TM, Turtle L, Seaton RA, Wootton DG, et al. Co-infections, secondary infections, and antimicrobial use in patients hospitalised with COVID-19 during the first pandemic wave from the ISARIC WHO CCP-UK study: a multicentre, prospective cohort study. Lancet Microbe. 2021;2:e354-e65.

[4] Davidson AD, Williamson MK, Lewis S, Shoemark D, Carroll MW, Heesom KJ, et al. Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. Genome Med. 2020;12:68.

[5] Meyer B, Chiaravalli J, Gellenoncourt S, Brownridge P, Bryne DP, Daly LA, et al. Characterising proteolysis during SARS-CoV-2 infection identifies viral cleavage sites and cellular targets with therapeutic potential. Nat Commun. 2021;12:5553.

[6] Hiscox JA, Mawditt KL, Cavanagh D, Britton P. Investigation of the control of coronavirus subgenomic mRNA transcription by using T7-generated negative-sense RNA transcripts. J Virol. 1995;69:6219-27.

[7] Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill. 2020;25.

[8] Dong X, Penrice-Randal R, Goldswain H, Prince T, Randle N, Donovan-Banfield I, et al. Analysis of SARS-CoV-2 known and novel subgenomic mRNAs in cell culture, animal model,

and clinical samples using LeTRS, a bioinformatic tool to identify unique sequence identifiers. Gigascience. 2022;11.

[9] Parker MD, Lindsey BB, Leary S, Gaudieri S, Chopra A, Wyles M, et al. Subgenomic RNA identification in SARS-CoV-2 genomic sequencing data. Genome Res. 2021;31:645-58.

[10] Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2 Transcriptome. Cell. 2020;181:914-21 e10.

[11] Belizario J, Garay-Malpartida M, Faintuch J. Lung microbiome and origins of the respiratory diseases. Curr Res Immunol. 2023;4:100065.

[12] Dickson RP, Singer BH, Newstead MW, Falkowski NR, Erb-Downward JR, Standiford TJ, et al. Enrichment of the lung microbiome with gut bacteria in sepsis and the acute respiratory distress syndrome. Nat Microbiol. 2016;1:16113.

[13] Aljabr W, Alruwaili M, Penrice-Randal R, Alrezaihi A, Harrison AJ, Ryan Y, et al. Amplicon and Metagenomic Analysis of Middle East Respiratory Syndrome (MERS) Coronavirus and the Microbiome in Patients with Severe MERS. mSphere. 2021;6:e0021921.

[14] Ke S, Weiss ST, Liu YY. Dissecting the role of the human microbiome in COVID-19 via metagenome-assembled genomes. Nat Commun. 2022;13:5235.

[15] Liu J, Liu S, Zhang Z, Lee X, Wu W, Huang Z, et al. Association between the nasopharyngeal microbiome and metabolome in patients with COVID-19. Synth Syst Biotechnol. 2021;6:135-43.

[16] Candel S, Tyrkalska SD, Alvarez-Santacruz C, Mulero V. The nasopharyngeal microbiome in COVID-19. Emerg Microbes Infect. 2023;12:e2165970.

[17] Nardelli C, Scaglione GL, Testa D, Setaro M, Russo F, Di Domenico C, et al. Nasal Microbiome in COVID-19: A Potential Role of Corynebacterium in Anosmia. Curr Microbiol. 2022;80:53.

[18] Hoque MN, Sarkar MMH, Rahman MS, Akter S, Banu TA, Goswami B, et al. SARS-CoV-2 infection reduces human nasopharyngeal commensal microbiome with inclusion of pathobionts. Sci Rep. 2021;11:24042.

[19] Hernandez-Teran A, Mejia-Nepomuceno F, Herrera MT, Barreto O, Garcia E, Castillejos M, et al. Dysbiosis and structural disruption of the respiratory microbiota in COVID-19 patients with severe and fatal outcomes. Sci Rep. 2021;11:21297.

[20] Wu J, Liu W, Zhu L, Li N, Luo G, Gu M, et al. Dysbiosis of oropharyngeal microbiome and antibiotic resistance in hospitalized COVID-19 patients. J Med Virol. 2023;95:e28727.

[21] Battaglini D, Robba C, Fedele A, Tranca S, Sukkar SG, Di Pilato V, et al. The Role of Dysbiosis in Critically Ill Patients With COVID-19 and Acute Respiratory Distress Syndrome. Front Med (Lausanne). 2021;8:671714.

[22] Carroll MW, Haldenby S, Rickett NY, Palyi B, Garcia-Dorival I, Liu X, et al. Deep Sequencing of RNA from Blood and Oral Swab Samples Reveals the Presence of Nucleic Acid from a Number of Pathogens in Patients with Acute Ebola Virus Disease and Is Consistent with Bacterial Translocation across the Gut. mSphere. 2017;2.

[23] Kafetzopoulou LE, Efthymiadis K, Lewandowski K, Crook A, Carter D, Osborne J, et al. Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. Euro Surveill. 2018;23.

[24] Kafetzopoulou LE, Pullan ST, Lemey P, Suchard MA, Ehichioya DU, Pahlmann M, et al. Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. Science. 2019;363:74-7.

[25] Lysholm F, Wetterbom A, Lindau C, Darban H, Bjerkner A, Fahlander K, et al. Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. PLoS One. 2012;7:e30875.

[26] Prince T, Dong X, Penrice-Randal R, Randle N, Hartley C, Goldswain H, et al. Analysis of SARS-CoV-2 in Nasopharyngeal Samples from Patients with COVID-19 Illustrates Population Variation and Diverse Phenotypes, Placing the Growth Properties of Variants of Concern in Context with Other Lineages. mSphere. 2022;7:e0091321.

[27] Nasir JA, Kozak RA, Aftanas P, Raphenya AR, Smith KM, Maguire F, et al. A Comparison of Whole Genome Sequencing of SARS-CoV-2 Using Amplicon-Based Sequencing, Random Hexamers, and Bait Capture. Viruses. 2020;12:895.

[28] Moore SC, Penrice-Randal R, Alruwaili M, Randle N, Armstrong S, Hartley C, et al. Amplicon-Based Detection and Sequencing of SARS-CoV-2 in Nasopharyngeal Swabs from Patients With COVID-19 and Identification of Deletions in the Viral Genome That Encode Proteins Involved in Interferon Antagonism. Viruses. 2020;12.

[29] De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34:2666-9.

[30] Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:257-.

[31] Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:giab008.

[32] O'Toole A, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol. 2021;7:veab064.

[33] Lu J, Rincon N, Wood DE, Breitwieser FP, Pockrandt C, Langmead B, et al. Metagenome analysis using the Kraken software suite. Nat Protoc. 2022;17:2815-39.

[34] de Castilhos J, Zamir E, Hippchen T, Rohrbach R, Schmidt S, Hengler S, et al. Severe Dysbiosis and Specific Haemophilus and Neisseria Signatures as Hallmarks of the Oropharyngeal Microbiome in Critically Ill Coronavirus Disease 2019 (COVID-19) Patients. Clin Infect Dis. 2022;75:e1063-e71.

[35] Brueggemann AB, Jansen van Rensburg MJ, Shaw D, McCarthy ND, Jolley KA, Maiden MCJ, et al. Changes in the incidence of invasive disease due to Streptococcus pneumoniae, Haemophilus influenzae, and Neisseria meningitidis during the COVID-19 pandemic in 26 countries and territories in the Invasive Respiratory Infection Surveillance Initiative: a prospective analysis of surveillance data. Lancet Digit Health. 2021;3:e360-e70.

[36] Ling L, Lai CKC, Lui G, Yeung ACM, Chan HC, Cheuk CHS, et al. Characterization of upper airway microbiome across severity of COVID-19 during hospitalization and treatment. Front Cell Infect Microbiol. 2023;13:1205401.

Table 1. PANGO assigned lineages of SARS-CoV-2 in cell culture and clinical samples using sequencing data generated by the four different methodologies. The sample ID is the name given to either the cell culture (CC) or clinical sample (NPS: Nasopharyngeal Swab) used in this study and does reflect the original name of the sample. The method indicates which sequencing approach was used: SISPA (S), SISPA-L (SL), SISPA-ACTIC-L (SAL) and SISPA-RSLA-L (SRL). The lineage assignments were determined using Pangolin, and the 'Ambiguity score' provides an indication of the quality of the SARS-CoV-2 sequence and its lineage assignment. A score closer to 1 or 1 indicates a high level of confidence in the lineage assignment, while a score of zero suggests that sequence information had to be imputed. Samples assigned from the 'Designation hash' are considered high-confidence assignments, while 'Failed' indicates that sequence information did not meet the necessary criteria for lineage assignment.

Sample ID	**Method**	**Lineage**	**Ambiguity score**	**Note**
01 CC Ct (11)	S	B.1.238	0.9997	
01 CC Ct (11)	SL	B.1.238		Designation hash
Sample ID	**Method**	**Lineage**	**Ambiguity score**	**Note**
01 CC Ct (11)	S	B.1.238	0.9997	
01 CC Ct (11)	SL	B.1.238		Designation hash
01 CC Ct (11)	SAL	B.1.238	1	
01 CC Ct (11)	SRL	B.1.238	1	
02 CC Ct (13)	S	B.1.238	0.9995	
02 CC Ct (13)	SL	B.1.238	1	
02 CC Ct (13)	SAL	B.1.238	0.9997	
02 CC Ct (13)	SRL	B.1.238	1	
03 CC Ct (14)	S	B.1.238	0.9992	
03 CC Ct (14)	SL	B.1.238		Designation hash
03 CC Ct (14)	SAL	B.1.238	1	
03 CC Ct (14)	SRL	B.1.238		Designation hash

04 NPS Ct (22)	S	B.4	0.9997	
04 NPS Ct (22)	SL	B.4	0.9992	
04 NPS Ct (22)	SAL	B.4	1	
04 NPS Ct (22)	SRL	B.4	0.9988	
05 NPS Ct (24)	S	B.1	0.9995	
05 NPS Ct (24)	SL	B.1	0.9995	
05 NPS Ct (24)	SAL	B.1	0.9988	
05 NPS Ct (24)	SRL	B.1	0.9766	
06 NPS Ct (26)	S	B	0.7099	
06 NPS Ct (26)	SL	B.4	0.5363	
06 NPS Ct (26)	SAL	B.4	0.8844	
06 NPS Ct (26)	SRL	B	0.9130	
07 NPS Ct (26)	S	B.1	0.9804	
07 NPS Ct (26)	SL	B.1	0.9978	
07 NPS Ct (26)	SAL	B.1	0.9924	
07 NPS Ct (26)	SRL	B.1.238	0.9314	
08 NPS Ct (27)	S	B.1	0.9969	
08 NPS Ct (27)	SL	B.1	0.9931	
08 NPS Ct (27)	SAL	B.1.238	0.9922	
08 NPS Ct (27)	SRL	B.1.238	0.9934	
09 NPS Ct (27)	S	B.1	0.8628	
09 NPS Ct (27)	SL	B.1.12	0.9971	
09 NPS Ct (27)	SAL	B.1.12	0.9879	
09 NPS Ct (27)	SRL	B.1.397	0.9644	
10 NPS Ct (31)	S			Failed
10 NPS Ct (31)	SL			Failed
10 NPS Ct (31)	SAL	B.1	0.5884	
10 NPS Ct (31)	SRL			Failed
11 NPS Ct (33)	S	B.1.238	0.9976	
11 NPS Ct (33)	SL	B.1.238	0.9826	
11 NPS Ct (33)	SAL	B.1.238	0.9870	
11 NPS Ct (33)	SRL	B.1.238	0.9953	
12 NPS Ct (36)	S	B.1.238	0.9879	
12 NPS Ct (36)	SL	B.1.238	0.9934	
12 NPS Ct (36)	SAL	B.1.238	0.9893	

12 NPS Ct (36)	SRL	B.1.238	0.9879	

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: