A metagenomics workflow for SARS-CoV-2 identification, co-pathogen detection, and overall diversity

Daniel Castañeda-Mogollón, Claire Kamaliddin, Yan Liu, Abu Naser Mohon, Rehan Mujeeb Faridi, Faisal Khan, Dylan R. Pillai

Supplementary Methods

Sample collection

A total of 125 clinical NPS and TS samples were collected and tested by Alberta Precision Laboratories (APL) between March 2020 and February 2021 as per routine procedure; the remaining specimen was aliquoted and immediately frozen at -80°C until further analysis. SARS-CoV-2 testing at APL is performed using a RT-PCR laboratory developed test (LDT) as described previously targeting the Envelope (E) gene[1]. Patients were deemed to be symptomatic based on the presence of at least one of the following (not related to a known cause or pre-existing condition): a new onset or worsening cough, fever (>38°C), shortness of breath (continually out of breath or unable to breathe deeply), loss of smell and/or taste, sore throat, nasal congestion, conjunctivitis, or muscle/joint aches. Symptom screening was based on patient reporting to the sampling nurse using the standard APL procedure. Swabs were performed by trained personnel as part of the Alberta COVID-19 testing program. Remaining specimen material was aliquoted and immediately frozen at -80°C in the APL biorepository for research purposes. Nintety-six samples were taken from NPS (76.80%), and 29 were from TS (23.2%). A total of sixty-five samples of 125 (52%) were positive for SARS-CoV-2. Seventy-one samples were symptomatic (56.80%), and 54 came from asymptomatic individuals (43.20%). Symptom screening was based on patient reporting to the sampling nurse using the standard APL procedure.

Nucleic acid extraction

Samples were randomized in extraction batches including internal controls to assess the metagenome kitome. DNA and RNA were extracted using the Qiagen QIAamp® DNA Mini Kit (Cat. No./ID: 51306, Qiagen, Germany) and the Qiagen QIAamp® Viral RNA Mini Kit (Cat. No/ID 52906, Qiagen, USA) respectively. Both protocols were adapted from the manufacturer's protocol to optimize the extraction. For the DNA extraction we followed the manufacturer's protocol. We used 40 to 200 µL of sample was digested using 1 mg/mL lysozyme (BioScience, Ireland) for 10 m at 37°C to digest bacterial cell walls, followed by 10 m Proteinase K (0.4 mg) (Qiagen, Germany) and RNAse A (0.06 mg) digestion (Monarch, NEB, USA) at 56°C in AL buffer (Qiagen, Germany). Nucleic acid precipitation and silica-column purification was performed according to the manufacturer's instructions. Elution was modified to elute DNA in two subsequent centrifugation steps with 40 µL of nuclease-free water, and elutions were preserved on ice until further utilization within 24h.

RNA extraction was performed using the Qiagen QIAamp® Viral RNA Mini Kit (Cat. No/ID 52906, Qiagen, USA). Ninety to 120 µL of sample were treated with Proteinase K (0.4 mg, Qiagen, Germany) for 10 m at 56°C in AL lysis buffer (Qiagen, Germany). Lysates were precipitated with ethanol according to the manufacturer's instructions and loaded into silica spin

columns. Remaining DNA was removed by on-column DNAse digestion (5 µL of DNAse I, 5 µL of 0.09 M $MnCl_2$ (Z3188) and 40 µL Yellow Core Buffer (Z317C), part of Promega SV Total RNA Isolation System, Z3105, Promega Corporation, USA) for 15 m at room temperature. DNAse was inactivated using the DNAse stop solution provided in the Promega kit (Z312D, Promega Corporation, USA) before subsequent washing steps and elution according to the Qiagen kit manufacturer's instructions; a modification was made to elute the RNA in two centrifugation rounds of 40 µL of nuclease-free water. The DNA and RNA concentrations were measured using a Qubit Flex Fluorometer (ThermoFisher® Scientific) with the respective high-sensitivity (HS) DNA and RNA assay kits (ThermoFisher® Scientific Cat. No. Q33230 and Q32852, respectively).

cDNA Synthesis

Priming was performed in RNAse-free tubes using a 1:1 mix of random hexamers (NEB, MA, USA) and the spike enrichment primer mix in NEB first-strand synthesis buffer for 8 m at 94°C. First-strand synthesis was performed from the fragmented and primed RNA with NEBNext First-Strand Synthesis enzyme mix for 10 m at 25°C, 50 m at 42°C and 15 m at 70°C. Second strand synthesis was performed immediately using the NEBNext Ultra II non-directional second strand synthesis module (E6111, NEB, MA, USA) as per the manufacturer's recommendation. Final cDNA was purified using 1.8X volume of SPRIselect beads (Beckman Life Science, USA) on a magnetic rack (Permagen, USA). After two washes with 80% fresh ethanol, cDNA was eluted in 50 µL of nuclease-free water.

Internal controls, library preparation and sequencing

Five to 120 ng of samples (DNA and cDNA) in 50 µL of nuclease-free water were submitted to the sequencing facility in 96-well plates and immediately processed for library preparation. The samples were sheared to an average size of 350 bp with the Covaris ML230 acoustic sonicator (Covaris, USA). Libraries were prepared using the New England Biolabs NEBNext Ultra II Library Preparation for Illumina (E7645, NEB, MA, USA). Finally, the libraries were pooled and sequenced on a NovaSeq Illumina instrument (Illumina, USA) using a NovaSeq 300 cycle SP v1.5 kit set (Illumina, USA) for 2 x 150 bp paired-end sequencing with a theoretical output of 1.6 X 10^9 reads. Three independent runs were performed. For quality control, the PhiX phage was spiked at a final concentration of 1% to assess the sequencing error rate. Inter-run reproducibility was assessed using a quantified oral microbiome mix (MSA-2004™, ATCC, USA) by comparing the inter-run reads per million (rPM) attributed to each species.

DNA extraction efficacy was assessed using a commercially available oral microbiome whole cell mix (MSA-2004™, ATCC, USA) containing six bacterial species (*Schaalia odontolytica, Prevotella melaninogenica, Fusobacterium nucleatum subsp. nucleatum, Streptococcus mitis, Veillonella parvula* and *Haemophilus parainfluenzae*). RNA extraction and cDNA synthesis performance were evaluated with five intact inactivated RNA viruses (Influenza A H1N1 PDMO A/NY/02/09, Influenza B Florida/0206, rhinovirus 1A, human metapneumovirus 8 Peru62006, and human respiratory syncytial virus A2; obtained from NATtrol™ Pneumonia Verification Panels NATPPQ- ZeptoMetrix, USA). Each RNA virus control was individually extracted following the same procedure used for the clinical samples. Viral extracts were either pooled equally and subjected to cDNA synthesis (pre-cDNA) or were used for cDNA synthesis before pooling (post-cDNA) to assess bias introduced through cDNA synthesis, before being

sent for library preparation and sequencing. No poly-A enrichment step nor ribosomal depletion was carried out during the pipeline.

Metagenome description and identification of infectious agents

Low-quality reads (reads with >10% of uncalled bases and/or less than 0.98 accuracy), duplicates, and low-complexity reads were removed using Paired-Read Iterative Contig Extension (PRICE) [2] software in conjunction with CD-HIT [3] and a Lempel-Ziv-Welch compression-score filtering algorithm. Finally, the IDseq pipeline utilizes the NCBI nucleotide (nt) and non-redundant protein (nr) databases for taxonomy identification through GSNAPL and RAPsearch2. A non-parametric test was performed to determine significant species between clusters. A background model from the negative water controls was generated for each of the three sequencing runs. An alpha and beta diversity analysis were performed across each group. Briefly, the alpha diversity Shannon index was used to measure richness and eveness of the observed OTUs. The beta diversity analysis using PCoAs summarizes sequence dissimilarities between clusters by using principal coordinates (PCs), which explains a fraction of the data variability in a two dimensional plot [4]. In other words, a PCoA represents a multidimensional sequence variability plot into a two-dimensional graph. For the variability assessment, the Bray-Curtis metric was used to determine the distance between the points by using a quantitative DNA/cDNA read approach, and its complementary counterpart (Jaccard) was used to determine the distance between the points by a qualitative approach (not taking into account the number of reads).

SARS-CoV-2 genome assembly and variant calling

Samples with reads mapped to the SARS-CoV-2 genome were submitted to the IDseq pipeline for genome assembly and variant calling through iVar. SNPs were called for variation analysis and compared against the reference genome using the default parameters from IDseq. Brielfy, the IDseq pipeline employs the bcftools software for variant calling. A SNP or deletion is called for reads with a quality score of 20 or above, a minimum read depth of 10 nucleotides, and the frequency of the SNP must be 75% or higher. For sample lineage characterization, genomes with a minimum breadth of coverage of 50% were submitted to the Pangolin online sequence aligner[5] (based on the GISAID consortium https://www.gisaid.org/ – available sequences on March 27th, 2021). A multiple sequence alignment (MSA) file was generated using Multiple Alignment based on Fast Fourier Transform (MAFFT v7.475; --thread 180 -- auto) [6]. The MSA file was used to generate a phylogenomics tree with IQtree (B 10000 –T AUTO –m MFP) [7]. Nodes with a bootstrap value below 0.50 were collapsed to the next closest node. Lineage characterization for samples between 25% to 50% breadth of coverage were estimated by the closest clade in the phylogenomics tree.

Identification of putative respiratory pathogens

The following species were screened as part of a putative respiratory pathogen panel*: Acinetobacter baumannii, Adenovirus, Aspergillus fumigatus, Blastomyces dermatitidis,* Bordetella (*Bordetella pertussis, Bordetella parapertussis, Bordetella bronchiseptica*), *Burkholderia multivorans, Chlamydia pneumoniae, Coccidioides imitis, Cryptococcus neoformans, Dolosigranulum pigrum,* Enterobacter (*E. asburiae, E. cloacae, E. hormaechei, E.*

sichuanensis, E. sp. R4-368, E. roggenkampii), *Haemophilus influenzae, Haemophilus parainfluenzae, HHV-6, histoplasma capsulatum, Human betaherpes virus6, Human coronavirus HKU1, Human coronavirus NL63, Human coronavirus 229E, Human coronavirus OC43, Influenza A virus, Influenza A/H1 virus, Influenza A/H3 virus, influenza A/H1-2009 virus, Klebsiella pneumoniae, Moraxella catharralis, Morbillivirus, Mucor spp., Mycobacteria tuberculosis, Mycoplasma pneumoniae, Parainfluenza virus 1, Parainfluenza virus 2, Parainfluenza virus 3, Parainfluenza virus 4, Parvovirus, Pseudomonas aeruginosa, Respiratory syncytial virus, Respirovirus, Rhinovirus A, Rhizopus spp., Rubulavirus, Serratia marcescens, Staphylococcus aureus, Streptococcus pyogenes and Streptococcus pneumoniae).*

Statistical analysis

A Wilcoxon non-parametric two-sided test was used to determine significance in the number of non-human reads in the DNA and RNA-metagenome between anatomical sampling sites and by SARS-CoV-2 diagnosis. Similarly, a Wilcoxon test was carried out to determine significance in the number of human reads in in the DNA and RNA-metagenome analysis between anatomical sampling sites. Principal coordinate analysis (PCoA) plots were performed for the DNA-metagenome in all taxonomic ranks, and in the RNA-metagenome for bacteriophages. The PCoA plots were generated using the Bray-Curtis and Jaccard distance matrix. A permutational multivariate analysis of variance (PERMANOVA) was performed for all PCoA plots to determine any significant clusters. Parallelly, the alpha-diversity Shannon index was calculated. To determine if the Ct value plays a key role in the non-significant findings by the diversity analysis, SARS-CoV-2-positive samples with a Ct value of 30 or above were excluded.

A Wilcoxon non-parametric two-sided test with Benjamini-Hochberg *p*-value correction was performed for all the identified species grouped by patient SARS-CoV-2 diagnosis, presence of symptoms, and anatomical sampling site. Exponential regression models were generated to determine any significant correlation between the number of SARS-CoV-2 mapped reads to the corresponding Ct-value of the samples. Pearson, Spearman, and R^2 values were generated as measures of correlation. Receiver operating characteristic (ROC) curves were generated to determine the sensitivity and specificity of the metagenomics pipeline for determining the presence of SARS-CoV-2. E-gene RT-PCR, performed as previously discussed, was employed as the gold standard test for determining clinical positivity.

A Chi-square goodness of fit test was performed to determine if all SNPs observed in SARS-CoV-2 genes were significantly different from the expected number of SNPs. The distance matrices, PERMANOVA, Shannon index calculation and PCoA plots were generated using the QIIME 2 bioinformatics platform [8]. The non-parametric statistical analysis, regression models, ROC curves, and plots were generated using GraphPad Prism version 8.4.3 for Mac [9]. The online iTOL software (v 6.1) and the iTOL annotation editor (v.1.4) were used for tree visualization[10]. All statistical tests were considered significant for *p*-values below 0.05. All analysis and bioinformatics processes were carried out in the Amazon Elastic Computing Cloud (EC2) from Amazon Web Services (AWS, Amazon, Seattle, USA) in a c5.24xlarge instance (17 TB of storage; 96 vCPUs, 192 GB of memory). Figure 1 was prepared with BioRender (biorender.com, Canada).

References

1. Pabbaraju, K. *et al.* Development and validation of RT-PCR assays for testing for SARS-CoV-2. *Official Journal of the Association of Medical Microbiology and Infectious Disease Canada* e20200026 (2021) doi:10.3138/jammi-2020-0026.
2. Jg, R., P, B. & Jl, D. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3 (Bethesda)* **3**, 865–880 (2013).
3. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
4. Goodrich, J. K. *et al.* Conducting a Microbiome Study. *Cell* **158**, 250–262 (2014).
5. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *bioRxiv* 2020.04.17.046086 (2020) doi:10.1101/2020.04.17.046086.
6. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–3066 (2002).
7. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268–274 (2015).
8. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* **37**, 852–857 (2019).
9. Home - GraphPad. https://www.graphpad.com/.
10. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **47**, W256–W259 (2019).