**Title: Use of whole genome sequencing to investigate a cluster of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) infections in emergency department personnel**

**Authors**: Ernest R. Chan PhD1\*, Lucas D. Jones BS2\*, Sarah N. Redmond BS3, Maria E. Navas MD4, Nataliya M. Kachaluba BS4, Trina F. Zabarsky CNP5, Davinder Bhullar MD6, Jennifer L. Cadnum, BS7, Peter A. Zimmerman PhD8α and Curtis J. Donskey MD9 α

**Supplementary file**

***SARS-CoV-2 sequencing and data analysis***

The amplicon-based next-generation sequencing (NGS) CleanPlex® SARS-CoV-2 Panel (Paragon Genomics) was used to sequence the whole genome of SARS-CoV-2 on an Illumina MiSeq sequencing platform. Sequences were aligned to the reference COVID-19 sequence (NCBI accession: NC\_045512.2) using Bowtie2.1 Variant positions were defined as any genomic position that had an alternate allele frequency >50% in any of the sequenced samples and were determined using samtools and mpileup.2-3 The reference allele frequency was determined at each base position by the count of reads with the Wuhan reference allele call divided by the total reads covering that base for each sample. Sequences were assigned to a transmission cluster if they had <2 base differences based on their variant positions.

Eight of the 11 samples had 100% of the genome covered by at least 100X coverage and the remaining 3 had 99% genome coverage at that depth. In total, 40 variant positions (reference allele frequency <50%) were identified. The range in reference allele frequencies at the variant positions demonstrate the utility of the deep sequencing approach and the ability to detect emerging mutations and minor strain differences.

Consensus sequences were generated using the major allele frequency and NC\_045512.2 as the reference sequence.2-3 Consensus sequences are available from GenBank and have been added to the Global Initiative on Sharing Avian Influenza Data (GISAID) hCoV-19 data portal. The consensus sequences for each strain was classified into 3 different commonly referenced clade classification systems using Nextstrain (nextstrain.org) and Pangolin (github.com/cov-lineages/pangolin). The supplementary table shows base changes for the sequenced SARS-CoV-2 samples in comparison to the Wuhan reference strain.

**Supplementary References**

1. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 2012;9:357-359.
2. Li H, Handsaker B, Wysoker A, et al. and 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 2009;25:2078-9.
3. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;27:2987-93.

Table. Base changes for sequenced SARS-CoV-2 sample in comparison to the Wuhan reference strain



\*, cluster 1; \*\*, cluster 2