**ONLINE DATA SUPPLEMENT for:**

# Viral Infection in Acute Exacerbation of Idiopathic Pulmonary Fibrosis

Sharon Chao Wootton, Dong Soon Kim, Yasuhiro Kondoh, Eunice Chen, Joyce S Lee, Jin Woo Song, Jin Won Huh, Hiroyuki Taniguchi, Charles Chiu, Homer Boushey, Lisa H. Lancaster, Paul J Wolters, Joseph DeRisi, Don Ganem, and Harold R Collard

**Deep sequencing library preparation**

Libraries were prepared as previously described (E1). In brief, nucleic acid extracted from BAL was first primed using abbreviated Illumina A and B adaptors attached to a unique 3-bp sequence tag (barcode) followed by a random hexamer. The 3-bp barcode was incorporated into each of the twelve samples to allow for 12-plex barcoded sequencing within a single lane on the Illumina flow cell. cDNA was amplified for 25 cycles of PCR using the barcoded adaptors, and the PCR product was run on a 4% native polyacrylamide gel at 4C to select for a narrow size distribution centered around 250-bp. The amplicons were then precipitated with 100% ethanol at 4C and resuspended in 16 microliters of water. Two microliters were carried into a second round of PCR amplification using the abbreviated A adaptor and a full length B adaptor for 15 cycles using 22-bp of the 3' end of the Illumina A adaptor and 61-bp of the Illumina B adaptor as primers. The product was size selected once more for products around 304-bp, which would carry the correct A/B topology. Ethanol precipitated DNA was then PCR amplified for ten cycles using the full length Illumina A adaptor and the 5' end of the B adaptor. This final library was sequenced on one lane of a paired end deep sequencing run with 65-bp read from each end of the insert.

**Deep sequencing analysis**

Low complexity reads with inadequate Lempel-Ziv-Welch (LZW) compression ratios were removed, and barcodes indicated by the sequence of the first 3-bp of each read were used to bin reads according to their original sample (E2). These reads were first filtered for host sequence through a high stringency BLAT to the human genome and transcriptome (E3), then screened for the presence of non-human sequence through iterative BLAST analysis to the NCBI NT database

(E4). First, high identity hits were isolated using a MEGABLAST with a word size of 28 against NT. Reads that did not align significantly to NT with a high word size were then aligned to NT again using MEGABLAST with word size of 12, followed by a sensitive BLASTN alignment with a word size of 7 and an e-value of 1e-3. All hits at every step were sorted according to NCBI taxonomy identifiers.

**ONLINE DATA SUPPLEMENT REFERENCES**

(E1) Yozwiak NL, Skewes-Cox P, Gordon A, Saborio S, Kuan G, Balmaseda A, Ganem D, Harris E, DeRisi JL. Human enterovirus 109: a novel interspecies recombinant enterovirus isolated from a case of acute pediatric respiratory illness in Nicaragua. *J Virol* 2010;84:9047-9058.

(E2) Welch TA. A technique for high-performance data compression. *IEEE Computer* 1984;17:9-19.

(E3) Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res* 2002;12:656-664.

(E4) Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389-3402.