Use of Unamplified RNA/cDNA–Hybrid Nanopore Sequencing for Rapid Detection and Characterization of RNA Viruses

Andy Kilianski, Pierce A. Roth, Alvin T. Liem, Jessica M. Hill, Kristen L. Willis, Rebecca D. Rossmaier, Andrew V. Marinich, Michele N. Maughan, Mark A. Karavis, Jens H. Kuhn, Anna N. Honko, C. Nicole Rosenzweig

Nanopore sequencing, a novel genomics technology, has potential applications for routine biosurveillance, clinical diagnosis, and outbreak investigation of virus infections. Using rapid sequencing of unamplified RNA/cDNA hybrids, we identified Venezuelan equine encephalitis virus and Ebola virus in 3 hours from sample receipt to data acquisition, demonstrating a fieldable technique for RNA virus characterization.

Portable and reliable molecular epidemiology techniques and field approaches for assessing virus genomes are desired to inform clinical diagnostics and public health operations. Need for such methods has been highlighted by the recent Middle East respiratory syndrome and Ebola virus disease (EVD) epidemics, during which it became necessary to characterize novel viruses and to evaluate genetic drift, transmission chains, and zoonotic introductions.

To determine if nanopore sequencing can be used as an accelerated viral genome sequencing tool, we utilized a rapid cDNA/RNA–hybrid library preparation procedure to sequence cell cultures of Venezuelan equine encephalitis virus vaccine (VEEV) strain TC-83 or Ebola virus (EBOV) isolate Makona-C05 stock IRF0137. To evaluate nanopore sequencing for rapid, field-deployable pathogen characterization, we collected raw read data and statistics for VEEV and EBOV sequence runs on the MinION sequencing device (Oxford Nanopore Technologies, Oxford, UK). To determine the level of identification and accuracy of genome characterization over sequencing runtime, these reads were then mapped to VEEV and EBOV genomes and to reference databases (RefSeq [http://www.ncbi.nlm.nih.gov/RefSeq/]). From the results of these analyses, we determined that the current and future versions of nanopore sequencing technology can be used to rapidly identify and characterize pathogens.

Author affiliations: US Army Edgewood Chemical Biological Center, Aberdeen Proving Ground, Maryland, USA (A. Kilianski, P.A. Roth, A.T. Liem, J.M. Hill, K.L. Willis, R.D. Rossmaier, A.V. Marinich, M.N. Maughan, M.A. Karavis, C.N. Rosenzweig); Defense Threat Reduction Agency, Fort Belvoir, Virginia, USA (K.L. Willis); National Institutes of Health, Fort Detrick, Frederick, Maryland, USA (J.H. Kuhn, A.N. Honko)

DOI: http://dx.doi.org/10.3201/eid2208.160270

The Study

This approach for pathogen identification and characterization differs from the previously used methods on the MinION platform. Biased techniques, such as amplicon sequencing, have proven to be effective in complex sample backgrounds in which titers of the target pathogen might be low, but such approaches limit characterization to known pathogens and require additional viral genome amplification (*1*–*4*). Unbiased techniques that require viral genome amplification (*5*) or that have been optimized for bacterial genomes (*6*,*7*) require longer sample and library preparation times, but can detect low pathogen titers or create highly accurate genomic data. We sequenced unamplified poly(A)-tailed viral RNA using rapid cDNA library preparation coupled with real-time data analysis to determine its potential application for pathogen genomic characterization.

VEEV has a single-stranded, linear, poly(A)-tailed RNA genome. Thus, poly-dT primers can be used for cDNA production without further genomic RNA manipulation. The workflow to isolate the RNA and prepare it for sequencing (online Technical Appendix, http://wwwnc.cdc.gov/ article/22/8/16-0270-Techapp1.pdf) took ≈3 hours from the initiation of sample processing to data acquisition on MinION (Figure, panel A). The sequencing of VEEV genomic RNA/cDNA hybrids attained in hours by using MinION revealed reads that mapped to the VEEV TC-83 genome within minutes by using the LAST (Computational Biology Research Consortium, Tokyo, Japan) multiple sequence alignment program (online Technical Appendix; Figure, panel B [*2*,*4*,*6*]). The coverage increased from 15–60 min from the 3′ end of the VEEV genome with reads aligning directionally from the 3′ to 5′ end of the VEEV genome (Figure, panel B). These alignment characteristics are indicative of the poly-dT priming strategy for poly(A)-tailed RNA.

To determine if the reads generated from VEEV TC-83 would align to the correct viral genome within a set

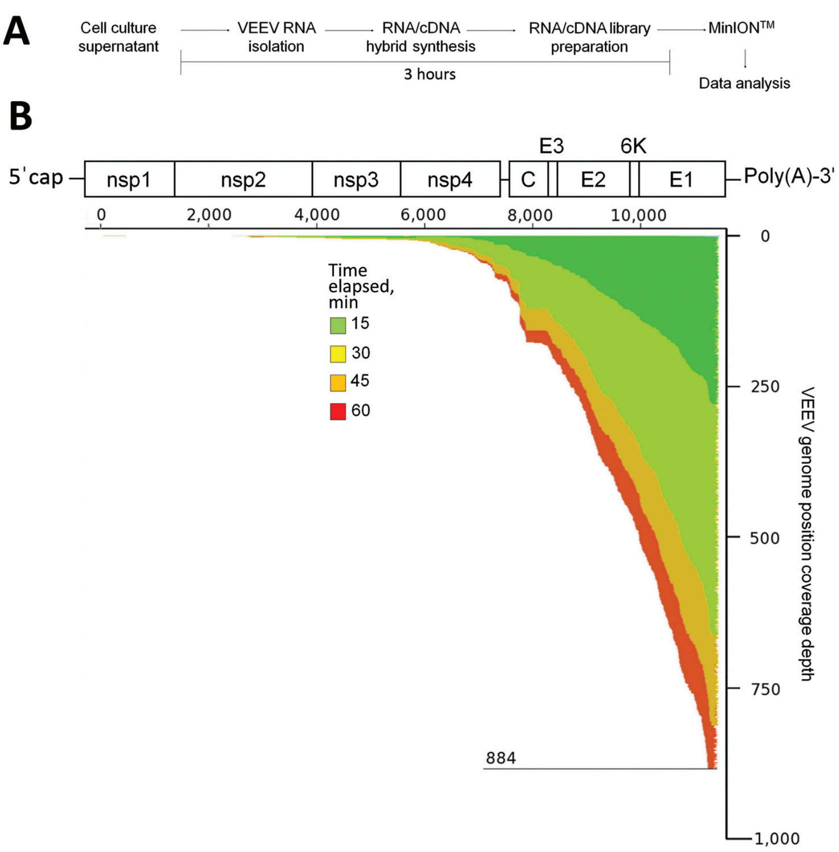

Figure. Use of unamplified RNA/cDNA--hybrid nanopore sequencing for genomic characterization of Venezuelan equine encephalitis virus (VEEV) TC-83. A) Sample preparation workflow for nanopore sequencing. First, viral RNA from BHK21 cell cultures of VEEV TC-83 was isolated, then single strand complimentary DNA (cDNA) was synthesized. The resulting RNA/cDNA hybrids were then prepared for nanopore sequencing and sequenced with data analysis occurring in real time. B) Genome organization and sequencing coverage over time of VEEV TC-83. VEEV is an alphavirus; its genome consists of a single strand of positive-sense RNA that can be translated into a polyprotein. Translation is critically dependent on the genomic 3′ poly(A)-tail. This tail can be used for reverse transcription priming by using poly-(dT) primers that anneal to it. Read data were aligned to VEEV TC-83 (accession number L01443) by using the multiple sequence alignment program LAST (Computational Biology Research Consortium, Tokyo, Japan [online Technical Appendix, wwwnc.cdc.gov/ article/22/8/16-0270-Techapp1.pdf). The coverage map shows the depth of genome coverage over 15, 30, 45, and 60 minutes of sequencing runtime, with the greatest depth observed at the 3′ end of the VEEV genome. Nsp, nonstructural protein; C, capsid; E, envelope.

of reference sequences, we used the viral genome reference sequences (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239&opt=Virus), plus the VEEV TC-83 genome database (GenBank accession no. L01443). We then used LAST to align nanopore reads against this set of references (online Technical Appendix). These alignments were used to generate a top hit table and the associated read alignment statistics against each hit. VEEV TC-83 was the top hit based on LAST alignment versus virus RefSeq genome sequences; wild-type VEEV placed second (Table). VEEV TC-83 was also identified as the top hit when the 15- and 60–min sequencing datasets were compared with alphavirus genome sequences (online Technical Appendix) (Table), demonstrating accuracy and depth achieved in a short time. We also analyzed the VEEV TC-83 dataset using the cloud-based metagenomic detection platforms Pathosphere (*8*) and One Codex (www.onecodex.com), and found that the sample contained VEEV (online Technical Appendix).

Molecular epidemiology, including use of viral genomics, played a major role during the 2013–2016 EVD response, informing contact tracing, diagnostic operability, and public health measures (*9*–*11*). To determine if EBOV is amenable to the same rapid sequencing methodology that was used for VEEV, unamplified negative-stranded RNA isolated from EBOV in Trizol (Thermofisher Scientific, http://www.thermofisher.com/us/en/home/brands/product-brand/trizol.html) was poly(A)-tailed, a single complementary strand of cDNA synthesized, and RNA/cDNA hybrids sequenced. The EBOV samples sequenced on MinION rapidly provided usable, accurate data, despite less raw data than the VEEV TC-83 dataset (137kbp for EBOV versus 2.4Mbp for VEEV at 60 min). Using 15-and 60-min time points and an identical alignment strategy to VEEV TC-83 above, we detected EBOV as the top hit within the sequencing dataset when compared to all virus RefSeq sequences (Table).

Despite success against the RefSeq database, the lack of depth within the dataset did not enable differentiation between the EBOV isolate sequenced here and the >1,500 EBOV draft genomes sequenced during the 2013–2016 outbreak (*10*,*12*,*13*), which indicates a limitation in this sequencing approach for negative-stranded RNA viruses. The poly(A)-tailing method was chosen because the reverse transcription primer adapters designed by Oxford Nanopore were developed to interact directly with the motor protein necessary for guiding DNA through the nanopores. This method greatly reduced preparation time and eliminated need for adaptor ligation reagents. This approach can

Table. Alignment statistics for detection of VEEV TC-83 and EBOV/Mak-C05 using unamplified RNA/cDNA-hybrid nanopore sequencing*

Virus samples and time points, min	Top hits (GenBank accession no.)	LAST score	Total bases mapped, %	Coverage, %	Average base depth	Per read accuracy, %
VEEV TC-83 (GenBank accession no. L01443)						
Viral genomes (RefSeq databases†)						
15	VEEV TC-83 (L01443)	138,321	5.54	76.14	50.94x	59–80
	VEEV WT (NC_001449.1)	789	0.05	18.59	1.76x	60–78
60	VEEV TC-83 (L01443)	419,153	17.17	78.54	153.16x	57–80
	VEEV WT (NC_001449.1)	1,182	0.08	32.12	1.82x	58–78
Alphavirus genomes						
15	VEEV TC-83 (L01443)	31,320	1.13	48.92	16.21x	67–69
	VEEV E541/73 (AF093102.1	6,463	0.27	95.07	5.26x	62–73
	VEEV 71–180 (AF069903.1)	5,865	0.22	30.18	5.08x	65–73
60	VEEV TC-83 (L01443)	96,348	3.55	48.92	50.84x	62–74
	VEEV E541/73 (AF093102.1)	21,411	0.89	99.91	16.36x	61–73
	VEEV 71–180 (AF069903.1)	16,429	0.64	51.04	8.78x	65–73
EBOV/Mak-C05 (GenBank accession no. KX000400)						
Viral genomes (RefSeq databases)						
15	EBOV/Mak-137 (KX000400)	529	0.11	9.29	1.00x	68
	Bovine herpesvirus (NC_024303.1)	73	0.02	0.18	1.00x	67
60	EBOV/Mak-137 (KX000400)	2,371	0.53	22.23	2.09x	66–71
	Bovine herpesvirus (NC_024303.1)	239	0.04	0.27	1.58x	67–74

*VEEV, Venezuelan equine encephalitis virus; EBOV, Ebola virus; LAST (Computational Biology Research Consortium, Tokyo, Japan), multiple sequence alignment program.
†RefSeq, NCBI Reference Sequence Database (http://www.ncbi.nlm.nih.gov/RefSeq/).

be revisited for sequencing negative-strand RNA viruses (*2*,*5*). Despite this limitation, the RefSeq alignments and nearest neighbor calls were possible with limited data, demonstrating the potential power of long-read rapid sequencing on nanopore platforms.

Conclusions

The current Middle East respiratory syndrome, EVD, and Zika virus disease outbreaks illustrate the necessity for rapid characterization of pathogens for environmental detection, clinical evaluation, and epidemiologic investigation. To determine whether nanopore sequencing can fill this role in a fieldable platform, we tested an RNA/cDNA–hybrid sequencing approach on VEEV TC-83 (a positive-stranded RNA virus) and EBOV (a negative-stranded RNA virus) prepared from cell-culture supernatants. This method definitively identified VEEV TC-83 and differentiated it from wild-type VEEV in ≈3 hours, including only 15 min of data acquisition on MinION. VEEV TC-83 was also differentiated from other alphavirus genomes, facilitating strain-level identification of TC-83. EBOV was also identified rapidly by this approach, differentiating the virus in the sample analyzed here from available virus reference genomes. However, variant/isolate level characterization was not possible due to limited data generated from the RNA/cDNA–hybrid approach.

The method applied here is greatly accelerated compared to traditional next-generation sequencing library preparation, and was used with reagents and equipment suitable for austere conditions (e.g., little need for cold chain, steps not requiring PCR). This study confirmed the possibility of accurate RNA virus genome characterization from RNA/cDNA hybrids by using limited sample manipulation, albeit from relatively pure samples. If samples derived directly from clinical matrices (e.g., blood, saliva) were used, this method would probably not support the necessary depth to characterize virus genomes unless the pathogen titer within these samples was high. As the depth of sequence data obtained from nanopore sequencing approaches continues to improve (*14*) and other pore types (such as RNA-specific sequencing pores) are integrated into commercial products, these unamplified techniques can transition from the laboratory to the field for more complex analysis.

Utilization of nanopore sequencing in Western Africa (*2*,*3*) has demonstrated potential for its use, and newly developed methods like this RNA/cDNA–hybrid approach can be integrated into fieldable protocols. For the emerging Zika virus, insufficiently high virus titers in clinical samples usually necessitates virus culture before genomic sequencing (*15*). Genomic Zika virus isolate characterization efforts would greatly benefit from the approaches outlined here, especially regarding materials needed for genomic library preparation and the time reduction for strain-level identification (*15*). By preparing and sequencing RNA/cDNA hybrids, the sample-to-answer time for RNA sequencing is greatly reduced, providing pathogen identification and characterization rapidly to inform future decision making.

Dr. Kilianski is a National Research Council fellow in the BioSciences Division at Edgewood Chemical Biological Center. His research focuses on biosurveillance, emerging viral pathogens, and the identification and characterization of novel agents that threaten today’s warfighter.

The authors acknowledge the efforts of the Sabeti Laboratory and Broad Institute of MIT and Harvard for providing the consensus sequence for EBOV/Mak-C05

This project was funded through the United States Army and Edgewood Chemical Biological Center 219 Innovative Project initiative to A.K. A.K. is also supported by the US National Academy of Science and the US Defense Threat Reduction Agency as a National Research Council fellow. A.K. is part of the Oxford Nanopore MinION Access Program, and has received free materials for this research. This work was funded in part through the prime contract of Battelle Memorial Institute, Columbus, Ohio, with the US National Institute of Allergy and Infectious Diseases under Contract no. HHSN272200700016I. A subcontractor to Battelle Memorial Institute who performed this work is: J.H.K., an employee of Tunnell Government Services, Inc, Bethesda, MD. Information in this report is cleared for public release and distribution is unlimited.

References

1. Kilianski A, Haas JL, Corriveau EJ, Liem AT, Willis KL, Kadavy DR, et al. Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. Gigascience. 2015;4:12. http://dx.doi.org/10.1186/s13742-015-0051-z
2. Hoenen T, Groseth A, Rosenke K, Fischer RJ, Hoenen A, Judson SD, et al. Nanopore sequencing as a rapidly deployable Ebola outbreak tool. Emerg Infect Dis. 2016;22:331–4; Epub ahead of print.
3. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530:228–32. http://dx.doi.org/10.1038/nature16996
4. Wang J, Moore NE, Deng Y-M, Eccles DA, Hall RJ. MinION nanopore sequencing of an influenza genome. Front Microbiol. 2015;6:766. http://dx.doi.org/10.3389/fmicb.2015.00766
5. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med. 2015;7:99. http://dx.doi.org/10.1186/s13073-015-0220-9
6. Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, et al. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella.* Genome Biol. 2015;16:114. http://dx.doi.org/10.1186/s13059-015-0677-2
7. Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, et al. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat Biotechnol. 2014;33:296–300; Epub ahead of print. http://dx.doi.org/10.1038/nbt.3103
8. Kilianski A, Carcel P, Yao S, Roth P, Schulte J, Donarum GB, et al. Pathosphere.org: pathogen detection and characterization through a web-based, open source informatics platform. BMC Bioinformatics. 2015;16:416. http://dx.doi.org/10.1186/s12859-015-0840-5
9. Mate SE, Kugelman JR, Nyenswah TG, Ladner JT, Wiley MR, Cordier-Lassalle T, et al. Molecular evidence of sexual transmission of Ebola virus. N Engl J Med. 2015;373:2448–54. http://dx.doi.org/26465384
10. Park DJ, Dudas G, Wohl S, Goba A, Whitmer SLM, Andersen KG, et al. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. Cell. 2015;161:1516–26. http://dx.doi.org/10.1016/j.cell.2015.06.007
11. Kugelman JR, Sanchez-Lockhart M, Andersen KG, Gire S, Park DJ, Sealfon R, et al. Evaluation of the potential impact of Ebola virus genomic drift on the efficacy of sequence-based candidate therapeutics. MBiol. 2015;6. http://dx.doi.org/10.1128/mBio.02227-14
12. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345:1369–72. http://dx.doi.org/10.1126/science.1259657
13. Carroll MW, Matthews DA, Hiscox JA, Elmore MJ, Pollakis G, Rambaut A, et al. Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. Nature. 2015;524:97–101. http://dx.doi.org/10.1038/nature14594
14. Ip CLC, Loose M, Tyson JR, de Cesare M, Brown BL, Jain M, et al. MinION analysis and reference consortium: phase 1 data release and analysis. F1000Research. 2015;4:1075. http://dx.doi.org/10.12688/f1000research.7201.1 PMID: 26834992
15. Faria NR. Azevedo R do S da S, Kraemer MUG, Souza R, Cunha MS, Hill SC, et al. Zika virus in the Americas: early epidemiological and genetic findings. Science. 2016;352:345–9. http://dx.doi.org/10.1126/science.aaf5036

Address for correspondence: C. Nicole Rosenzweig, BioDefense Branch, BioSciences Division, Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD 21010, USA; email: carolyn.n.rosenzweig.civ@mail.mil

DOI: http://dx.doi.org/10.3201/eid2208.160270

Use of Unamplified RNA/cDNA-Hybrid Nanopore Sequencing for Rapid Detection and Characterization of RNA Viruses

Technical Appendix

Viral Growth and RNA Isolation

To determine the ability of nanopore sequencing to provide rapid genomic data on RNA virus pathogens, a workflow was adopted and developed from cDNA sequencing protocols created by Oxford Nanopore Technologies (Oxford, UK) (MAP SEQ-002) (Figure 1, panel A). In brief, , Venezuelan equine encephalitis virus (VEEV) vaccine strain TC-83 and Ebola virus (EBOV) variant Makona isolate C05 stock IRF0137 (EBOV/Mak-C05) were grown and RNA isolated from clarified cell-culture supernatants. VEEV TC-83 was prepared from stocks derived from United States Army Venezuelan equine encephalitis virus TC-83 stocks. One MOI of VEEV TC-83 was adsorbed on Vero E6 monolayers for 2 hours. After 48 h of incubation in Minimum Essential Medium-α + 10% fetal bovine serum (GIBCO, Gaithersburg, MD; ThermoFisher Scientific, Pittsburgh, PA), cell culture supernatants were collected and clarified by centrifugation at 650 × *g* for 10 min at 4°C. RNA was isolated from cell-culture supernatant using the QIAamp MinElute virus spin kit (QIAGEN) for isolated VEEV particles.

The C05 isolate of the Makona variant of Ebola virus (full designation: Ebola virus/H.sapiens-tc/GIN/2014/Makona-C05, abbreviation: EBOV/Mak-C05) was isolated in 2014 in Vero E6 cells and kindly provided by Dr. Gary P. Kobinger (Public Health Agency of Canada, Winnipeg, Canada, BioSample: SAMN03611815, internal reference IRF0135). Vero E6 cells were used to propagate EBOV by two additional tissue culture passages in Vero E6 cells using Minimum Essential Medium-α, GlutaMAX, no nucleosides (GIBCO, ThermoFisher Scientific) supplemented with 2% US-origin, certified, heat-inactivated fetal bovine serum (HI-FBS, GIBCO, ThermoFisher Scientific). Following harvest, HI-FBS was QS'd to 10% final concentration before cryopreservation. GenBank accession no. KX000400 BioSample:

SAMN04490241, internal reference IRF0137. RNA was isolated from virus preps in 1:4 supernatant:Trizol after RNA extraction with cleanup using RNeasy MinElute cleanup kit (QIAGEN).

RNA/cDNA-Hybrid Sequencing Preparations

A total of 250 ng of VEEV or EBOV RNA were either directly added to a single-strand cDNA reaction (VEEV) primed using poly-dT primers provided by Oxford Nanopore Technologies (ONT; DEV-MAP003) or poly(A)-tailed (*Escherichia coli* poly(A) polymerase, (New England BioLabs) before addition (EBOV). cDNA synthesis was performed using SuperScript II reverse transcription 18064–014 (Life Technologies, Carlsbad, CA) per standard manufacturers protocols at 50°C for 50 min, followed by 70°C for 15 min. After cDNA synthesis, cDNA/RNA hybrids were prepared for nanopore sequencing (DEV-MAP003) by purifying RNA/cDNA hybrids using 0.7× Agencourt AmPure XP beads (Beckman Coulter, Fullerton, CA) followed by 2× 80% ethanol washes. Purified RNA/cDNA hybrids were then incubated with binding buffer for 45 min (ONT; DEV-MAP003), motor protein for 5 min (ONT; DEV-MAP003), then loading buffer for 5 min (ONT; DEV-MAP003) per the manufacturers protocol (ONT; DEV-MAP003). Prepared libraries were then diluted per manufacturers protocol with water and fuel mix (ONT; DEV-MAP003) to a final volume of 300 µl, with 150 µL sequencing solution loaded with p1000 tips onto individual MinION flow cells (7.3) for sequencing.

Data Collection and Analysis

Data were collected in real time using Oxford Nanopore software (VEEV: MinKNOW: 0.50.1.15, Metrichor: 1.13.1; EBOV: MinKNOW: 0.48.2.14, Metrichor: 1.10.1) and analyzed for genome alignment using LAST-648 (*1*) (lastal) with the options (-s 2 -T 0 -Q 0 -a 1). Duplications in the alignments were removed using last-map-probs (*1*). These methods are identical to ones used for previous work with amplicon and native nucleic acid sequencing (*2*–*6*). Nanopore reads were aligned against the Viral Genomes database (*7*) of viral reference genome sequences (data for complete genomes: Viruses (taxid 10239); 6,635 total entires, with the 2 stock genomes added from this study) and top hits and alignment statistics were generated.

VEEV reads were also aligned against a database of alphavirus genome sequences (1,440 total entries; detailed below). The read files, associated data and scripts are available at GenBank EU accession numbers SAMEA3865262 (VEEV) and SAMEA3865263 (EBOV) for nanopore data and NCBI BioProject PRJNA311755 and GenBank accession no. KX000400 for the EBOV reference. Sample use of Pathosphere (*8*) for analysis is publically available at www.pathosphere.org under the manuscript header. One Codex analysis (*9*) is also publically available at https://app.onecodex.com/analysis/public/57c08743784b41d3. All reference files used for the analysis above are available as 'viralRefSeq.fa' and 'all_alphavirus_complete.fasta' attached here. Scripts are available below in addition to being hosted at www.pathosphere.org.

References

1. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21(3):487–93. http://dx.doi.org/10.1101/gr.113985.110 **PMID: 21209072**

2. Hoenen T, Groseth A, Rosenke K, Fischer RJ, Hoenen A, Judson SD, et al. Nanopore sequencing as a rapidly deployable Ebola outbreak tool. Emerg Infect Dis. 2016;22; Epub ahead of print.

3. Laver TW, Caswell RC, Moore KA, Poschmann J, Johnson MB, Owens MM, et al. Pitfalls of haplotype phasing from amplicon-based long-read sequencing. Sci Rep. 2016;6:21746. http://dx.doi.org/10.1038/srep21746 **PMID: 26883533**

4. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, et al. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif. 2015;3:1–8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4691839/ **PMID: 26753127**

5. Wang J, Moore NE, Deng Y-M, Eccles DA, Hall RJ. MinION nanopore sequencing of an influenza genome. Front Microbiol. 2015;6:766. http://dx.doi.org/10.3389/fmicb.2015.00766 **PMID: 26347715**

6. Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, et al. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. Genome Biol. 2015;16:114. http://www.ncbi.nlm.nih.gov/pubmed/26025440 **PMID: 26025440**

7. Brister JR, Ako-adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. Nucleic Acids Res. 2014;43:D571–7. http://www.ncbi.nlm.nih.gov/pubmed/25428358 http://dx.doi.org/10.1093/nar/gku1207

8. Kilianski A, Carcel P, Yao S, Roth P, Schulte J, Donarum GB, et al. Pathosphere.org: pathogen detection and characterization through a web-based, open source informatics platform. BMC Bioinformatics. 2015;16:416. http://dx.doi.org/10.1186/s12859-015-0840-5 **PMID: 26714571**

9. Minot SS, Krumm N, Greenfield NB. One Codex: A sensitive and accurate data platform for genomic microbial identification. bioRxiv. 2015 Sep 28; Epub ahead of print. http://biorxiv.org/content/early/2015/09/28/027607

Scripts

```
#!/bin/bash # last_time_alignment_pipeline
# # This is the main script that aligns reads
using LAST according to timestamps #
produced by the ExtractTimesFromReads
program. It converts the fast5 files to a #
fastq, extracts only the reads that were
produced before specified time argument,
# prepares the reference database for last,
aligns and then outputs the top 10 reference
hits from
# the database.
#
# Dependencies that must be installed:
LAST, JAVA, perl, bash, HDF5 libraries.
#
# This program was developed exclusively
with government funds by
# OptiMetrics, Inc. in support of U.S. Army
Edgewood Chemical Biological
# Center.
#
# Copyright (C) 2016 OptiMetrics, Inc.
#
# This program is free software: you can
redistribute it and/or modify
# it under the terms of the GNU General
Public License as published by
# the Free Software Foundation, version 3 of
the License.
#
# This program is distributed in the hope
that it will be useful,
# but WITHOUT ANY WARRANTY;
without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR
A PARTICULAR PURPOSE. See the
# GNU General Public License for more
details.
#
# You should have received a copy of the
GNU General Public License
# along with this program. If not, see
<http://www.gnu.org/licenses/>.
#
set -o nounset;
set -o errexit;
args=($@);
len=${#args[@]};
min_scripts=”$(cd “$(dirname
“${BASH_SOURCE[0]}”)” && pwd)”;
if [ $len -lt 4 ] || [ $len -gt 4 ]
then
echo “Usage:
last_time_alignment_pipeline.sh <fast5
input directory> <reference fasta> <time in
seconds> <output directory>”
exit;
fi
fast5_dir=${args[0]};
reference=${args[1]};
time=${args[2]};
output_dir=${args[3]};
mkdir -p $output_dir;
cd $output_dir;
reads=$output_dir/reads.fq;
if [ ! -e $reads ]
then
echo “converting fast5 to fastq ....”;
ls $fast5_dir/*.fast5 | xargs -I [] java -jar -
Djava.library.path=$min_scripts/lib
```

```
$min_scripts/Fast5toFastq.jar [] >
$output_dir/reads.fq
fi
timestamps=$output_dir/timestamps.csv;
if [ ! -e $timestamps ]
then
echo “extracting timestamps from fast5 ....”;
ls $fast5_dir/*.fast5 | xargs -I [] java -
Djava.library.path=$min_scripts/lib -cp
$min_scripts/lib/:$min_scripts/source
extracttimesfromreads.ExtractTimesFromRe
ads [] > $timestamps;
fi
echo “creating last db reference ...”;
lastdb ref $reference;
echo “filtering reads file by time ...”;
perl $min_scripts/filter_reads_by_time.pl
$time $time $timestamps $reads
$output_dir/reads.$time.fna
echo “running last alignment ...”;
lastal -r1 -a1 -b1 -q1 -Q0 ref
$output_dir/reads.$time.fna >
$output_dir/align.$time.maf
echo “running last-map-probs”;
/common/bin/last-658/last-map-probs
$output_dir/align.$time.maf >
$output_dir/align.$time.nodups.maf
echo “getting top scores ...”;
perl
$min_scripts/findTopAlignmentScores.pl
$output_dir/align.$time.nodups.maf
$output_dir/topscores.$time.txt
echo “DONE!”;
#$Log$
/**
* Fast5toFastq
*
* This program extracts the fastq data from a
fast5 file that has been
* processed using the Metrichor analysis
from Oxford Nanopore.
*
* This program was developed exclusively
with government funds by
* OptiMetrics, Inc. in support of U.S. Army
Edgewood Chemical Biological
* Center.
*
* Copyright (C) 2016 OptiMetrics, Inc.
*
* This program is free software: you can
redistribute it and/or modify
* it under the terms of the GNU General
Public License as published by
* the Free Software Foundation, version 3 of
the License.
* This program is distributed in the hope
that it will be useful,
* but WITHOUT ANY WARRANTY;
without even the implied warranty of
* MERCHANTABILITY or FITNESS FOR
A PARTICULAR PURPOSE. See the
* GNU General Public License for more
details.
*
* You should have received a copy of the
GNU General Public License
* along with this program. If not, see
<http://www.gnu.org/licenses/>.
*
*/
package fast5tofastq;
import java.io.BufferedReader;
import java.io.File;
import java.io.FileReader;
import java.util.StringTokenizer;
import ncsa.hdf.object.h5.*; // include the
HDF5 object package
import ncsa.hdf.hdf5lib.*; // include the Java
HDF5 interface
*
*/
public class Fast5toFastq
{
public static void main(String[] argv)
{
if (argv.length != 1)
{
System.out.println(“Usage: Fast5toFastq
<fast5_filename>”);
System.out.println(“Output prints to
STDOUT”);
try
{
}
catch (Exception e)
{
}
return;
}
// create an H5File object
H5File h5file = new H5File(argv[0],
HDF5Constants.H5F_ACC_RDONLY);
try
{
BufferedReader reader = null;
String path =
Fast5toFastq.class.getProtectionDomain().ge
tCodeSource().getLocation().toURI().getPat
h();
File configFile = new
File(path.substring(0,path.lastIndexOf("/")+
1) + "fast5tofastq.conf");
try
{
reader = new BufferedReader(new
FileReader(configFile));
}
catch (Exception e)
{
System.out.println("Config file
fast5tofastq.conf must be in the same
location as the fast5tofastq jar file.");
return;
}
h5file.open();
while (reader.ready())
{
String fastqPath = reader.readLine();
// open file and retrieve the file structure
try
{
H5ScalarDS obj = (H5ScalarDS)
h5file.get(fastqPath);
String [] s = (String [])obj.read();
for(int i = 0; i < s.length; i++)
System.out.println(s[i]);
}
catch (Exception e)
{
// This path doesn't exist in the file. Do
nothing, and move on.
}
}
reader.close();
}
catch (Exception ex)
{
System.err.println(ex);
}
try { h5file.close(); }
catch (Exception ex) {}
}
}
/**
* ExtractTimesFromReads
*
* This program extracts the read names and
read times from a fast5 file
* that has been processed using the
Metrichor analysis from Oxford Nanopore.
*
* This program was developed exclusively
with government funds by
* OptiMetrics, Inc. in support of U.S. Army
Edgewood Chemical Biological
* Center.
*
* Copyright (C) 2016 OptiMetrics, Inc.
*
* This program is free software: you can
redistribute it and/or modify
* it under the terms of the GNU General
Public License as published by
* the Free Software Foundation, version 3 of
the License.
* This program is distributed in the hope
that it will be useful,
* but WITHOUT ANY WARRANTY;
without even the implied warranty of
* MERCHANTABILITY or FITNESS FOR
A PARTICULAR PURPOSE. See the
* GNU General Public License for more
details.
```

```
*
* You should have received a copy of the
GNU General Public License
* along with this program. If not, see
<http://www.gnu.org/licenses/>.
*
*/
package extracttimesfromreads;
import java.io.BufferedReader;
import java.io.File;
import java.io.FileReader;
import java.util.List;
import ncsa.hdf.object.h5.*; // include the
HDF5 object package
import ncsa.hdf.hdf5lib.*; // include the Java
HDF5 interface
import ncsa.hdf.object.Attribute;
public class ExtractTimesFromReads
{
public static void main(String[] argv)
{
if (argv.length != 1)
{
System.out.println(“Usage:
ExtractTimesFromReads
<fast5_filename>”);
System.out.println(“Output prints to
STDOUT”);
System.exit(0);
}
// create an H5File object
H5File h5file = new H5File(argv[0],
HDF5Constants.H5F_ACC_RDONLY);
String name=””;
try
{
BufferedReader reader = null;
String path =
ExtractTimesFromReads.class.getProtection
Domain().getCodeSource().getLocation().to
URI().getPath();
File configFile = new
File(path.substring(0,path.lastIndexOf(“/”)+
1) + “fast5tofastq.conf”);
try
{
reader = new BufferedReader(new
FileReader(configFile));
}
catch (Exception e)
{
System.out.println(“Config file
fast5tofastq.conf must be in the same
location as the fast5tofastq jar file.”);
return;
}
h5file.open();
while (reader.ready())
{
String fastqPath = reader.readLine();
// open file and retrieve the file structure
try
{
H5ScalarDS obj = (H5ScalarDS)
h5file.get(fastqPath);
String [] s = (String [])obj.read();
name = s[0].substring(0, s[0].indexOf('\n'));
}
catch (Exception e)
{
// This path doesn't exist in the file. Do
nothing, and move on.
}
}
reader.close();
H5Group obj = null;
for(int i = 0; i < 1000; i++)
{
String readString = “Read_”+ i;
//System.err.println(readString);
obj = (H5Group)
h5file.get(“Analyses/EventDetection_000/R
eads/”+readString);
if(obj !=null)
{
break;
}
}
if(obj == null)
{
//System.err.println(name);
throw new Exception(“Read does not
exist.”);
}
List<Attribute> list = obj.getMetadata();
long startTime= –1;
long duration= –1;
for(int i = 0; i < list.size(); i++)
{
if (list.get(i).toString().equals(“start_time”))
{
startTime =
((long[])list.get(i).getValue())[0];
//System.err.println(“start_time” + ” = ” +
startTime);
}
if (list.get(i).toString().equals(“duration”))
{
duration = ((long[])list.get(i).getValue())[0];
//System.err.println(“duration” + ” = ” +
duration);
}
}
obj = (H5Group)
h5file.get(“UniqueGlobalKey/channel_id”);
list = obj.getMetadata();
double samplingRate = –1;
for(int i = 0; i < list.size(); i++)
{
if
(list.get(i).toString().equals(“sampling_rate”
))
{
samplingRate =
((double[])list.get(i).getValue())[0];
//System.err.println(“sampling_rate” + ” = ”
+ samplingRate);
}
}
double time= ((double)(startTime +
duration))/samplingRate;
System.out.println(name + “,”+time);
}
catch (Exception ex)
{
System.err.println(name);
System.err.println(ex);
ex.printStackTrace();
}
try { h5file.close(); }
catch (Exception ex) {}
}
}
```

```
#!/usr/bin/perl
# Filter_reads_by_time
#
# This program will pull out reads that were
generated before the specified time
# by the specified range and output it to a
fasta file.
#
# This program was developed exclusively
with government funds by
# OptiMetrics, Inc. in support of U.S. Army
Edgewood Chemical Biological
# Center.
#
# Copyright (C) 2016 OptiMetrics, Inc.
#
# This program is free software: you can
redistribute it and/or modify
# it under the terms of the GNU General
Public License as published by
# the Free Software Foundation, version 3 of
the License.
#
# This program is distributed in the hope
that it will be useful,
# but WITHOUT ANY WARRANTY;
without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR
A PARTICULAR PURPOSE. See the
# GNU General Public License for more
details.
#
# You should have received a copy of the
GNU General Public License
# along with this program. If not, see
<http://www.gnu.org/licenses/>.
#
eval \'exec /usr/bin/perl -S $0 “$@”'
if 0; # not running under some shell
use strict;
use warnings;
my $len=scalar(@ARGV);
if ($len < 5 || $len > 6){
&printUsage();
}
my $time=$ARGV[0];
my $range=$ARGV[1];
my $timestamps_csv=$ARGV[2];
my $reads_input=$ARGV[3];
my $reads_output=$ARGV[4];
sub printUsage {
print STDOUT “Usage:
filter_sam_by_time.pl <time> <range>
<timestamps_csv> <sam_input>
<sam_output>\n”; exit(–1);
}
my %tstamps;
### Program runs here ###
open(TIME,”<$timestamps_csv”);
while(<TIME>){
my $ln=$_;
chomp $ln;
if($ln=~m/^([^\,]+)\,(.*)$/){
my $sid=$1;
my $tstamp=$2;
$sid=~s/\_strand//g;
if($tstamp<$time && $tstamp>=($time-
$range)){
$tstamps{$sid}=$tstamp;
}
}
} close TIME;
open(READSIN,”<$reads_input”);
open(READSOUT,”>$reads_output”);
while(<READSIN>){
my $ln=$_;
chomp $ln;
my $readName=$ln;
if(defined($tstamps{$readName})){
$ln=~s/^\@/\>/;
print READSOUT $ln.”\n”;
$ln=<READSIN>;
chomp $ln;
print READSOUT $ln.”\n”;
my $null=<READSIN>;
$null=<READSIN>;
}
}
close READSIN;
close READSOUT;
#!/usr/bin/perl
# findTopAlignmentScores
#
# This program pools the scores of each last
alignment by reference.
# It then picks the top 10 references based
on their scores.
#
# This program was developed exclusively
with government funds by
# OptiMetrics, Inc. in support of U.S. Army
Edgewood Chemical Biological
# Center.
#
# Copyright (C) 2016 OptiMetrics, Inc.
#
# This program is free software: you can
redistribute it and/or modify
# it under the terms of the GNU General
Public License as published by
# the Free Software Foundation, version 3 of
the License.
#
# This program is distributed in the hope
that it will be useful,
# but WITHOUT ANY WARRANTY;
without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR
A PARTICULAR PURPOSE. See the
# GNU General Public License for more
details.
#
# You should have received a copy of the
GNU General Public License
# along with this program. If not, see
<http://www.gnu.org/licenses/>.
#
eval \'exec /usr/bin/perl -S $0 “$@”'
if 0; # not running under some shell
use strict;
use warnings;
my $len=scalar(@ARGV);
```

```
if ($len < 2 || $len > 2){
&printUsage();
}
my $alignment=$ARGV[0];
my $output=$ARGV[1];
sub printUsage {
print STDOUT “Usage:
findTopAlignmentScores.pl <alignment>
<output>\n”; exit(−1);
}
open(FIN,”<$alignment”);
my $last_score=0;
my %scores;
my $got_score=0;
while(<FIN>){
my $line=$_;
chomp $line;
if($line=~m/^a score\=(\d+)/){
$last_score=$1;
$got_score=1;
}
elsif($line=~m/^s\ ([^\ ]+\)/ &&
$got_score){
$got_score=0;
my $gi=$1;
if(!defined($scores{$gi})){
$scores{$gi}=$last_score;
}else{
$scores{$gi}+=$last_score;
}
}
}
close FIN;
open(FOUT,”>$output”);
my @keys=
sort{$scores{$b}<=>$scores{$a}}
keys(%scores);
my @vals=@scores{@keys};
for(my $i=0; $i<10; $i++){
if(defined($keys[$i])){
print FOUT $keys[$i].” score\
“.$vals[$i].”\n”;
}
}
close FOUT;
```