W366–W371 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* *Published online 22 May 2020*
doi: 10.1093/nar/gkaa413

NanoSPC: a scalable, portable, cloud compatible viral nanopore metagenomic data processing pipeline

Yifei Xu[1,2,*,†], Fan Yang-Turner[1,2,†], Denis Volk[1,2] and Derrick Crook[1,2]

[1]Nuffield Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK and [2]NIHR Oxford Biomedical Research Centre, University of Oxford, UK

Received March 05, 2020; Revised April 22, 2020; Editorial Decision May 04, 2020; Accepted May 11, 2020

ABSTRACT

Metagenomic sequencing combined with Oxford Nanopore Technology has the potential to become a point-of-care test for infectious disease in public health and clinical settings, providing rapid diagnosis of infection, guiding individual patient management and treatment strategies, and informing infection prevention and control practices. However, publicly available, streamlined, and reproducible pipelines for analyzing Nanopore metagenomic sequencing data are still lacking. Here we introduce NanoSPC, a scalable, portable and cloud compatible pipeline for analyzing Nanopore sequencing data. NanoSPC can identify potentially pathogenic viruses and bacteria simultaneously to provide comprehensive characterization of individual samples. The pipeline can also detect single nucleotide variants and assemble high quality complete consensus genome sequences, permitting high-resolution inference of transmission. We implement NanoSPC using Nextflow manager within Docker images to allow reproducibility and portability of the analysis. Moreover, we deploy NanoSPC to our scalable pathogen pipeline platform, enabling elastic computing for high throughput Nanopore data on HPC cluster as well as multiple cloud platforms, such as Google Cloud, Amazon Elastic Computing Cloud, Microsoft Azure and OpenStack. Users could either access our web interface (https://nanospc.mmmoxford.uk) to run cloud-based analysis, monitor process, and visualize results, as well as download Docker images and run command line to analyse data locally.

INTRODUCTION

Oxford Nanopore Technology (ONT) is a third generation sequencing technology with the advantage of generating real-time, long read data with highly portable devices. Nanopore sequencing has been successfully applied in a broad range of research areas, including human genetics, cancer, microbiology, plant, and infectious diseases (1–5). On-site sequencing with ONT has enabled real-time surveillance and tracking of Ebola, Zika and Lassa epidemics (6–8).

Metagenomic sequencing has the capacity to detect all potential pathogens from individual clinical samples, and provide genomic information for comprehensive characterization of the pathogens, microbiome analyses, and investigation of epidemiology and transmission (9–11). Metagenomic sequencing with ONT has the potential to become a point-of-care test for infectious diseases in clinical and public health settings, providing rapid diagnosis of infection, guide individual patient management and treatment strategies, and informing infection prevention and control practices (12–14). Nanopore metagenomic sequencing of the novel coronavirus disease 2019 (COVID-19) that causes the ongoing pandemic in the world has provided critical and timely evidence for human-to-human transmission of this virus (15).

Nanopore sequencing data is characterized by high error rates (approximately 10% with state of art chemistry and basecalling algorithms) compared to next generation sequencing (approximately 0.1%) (16). A variety of bioinformatic tools have been developed to overcome such high error rates and improve data quality for each step associated with the analysis, including basecalling, reads mapping, variants calling, and genome assembly (17–21). However, publicly available streamlined and reproducible pipelines for analyzing Nanopore metagenomic sequencing data are still lacking, which consequently could impede its application, especially for users with minimum bioinformatics knowledge.

Here, we introduce NanoSPC, a scalable, portable and cloud compatible pipeline for analyzing Nanopore sequencing data. NanoSPC can identify potentially pathogenic viruses and bacteria simultaneously to provide comprehensive characterization of individual samples. The pipeline can also detect single nucleotide variants and assemble high quality complete consensus genome sequences, which

[*]To whom correspondence should be addressed. Email: yifei.xu@ndm.ox.ac.uk
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

permits inference of transmission with high resolution. We implement NanoSPC using Nextflow pipeline manager within Docker images to allow reproducibility and portability of the analysis. We deploy NanoSPC to our scalable pathogen pipeline platform, enabling elastic computing for high throughput data via HPC cluster as well as multiple commercial cloud platforms.

METHODS AND FUNCTIONALITIES

Data input and quality control

NanoSPC takes Nanopore sequencing reads of the fastq format and, optionally, raw signal data of the fast5 format as input (Figure 1). We have tested the pipeline with fastq reads produced by the R9.4 and R9.4.1 version flow cells, native and rapid barcoding kits. The acceptable multiread fast5 format should contain data pertaining to multiple reads in each fast5 file. While NanoSPC is designed to analyze Nanopore sequencing data, some of its modules could be used to analyze sequencing data from other platforms, such as PacBio or Illumina. The pipeline investigates the quality of the sequencing reads using NanoPlot (22). A comprehensive statistical summary is produced to report the overall data quality, including number of reads, total nucleotide bases, mean and median read length, and quality scores. In addition, a variety of informative graphs are generated to display multiple aspects of the data, such as cumulative yield plot showing efficiency of the flow cell against time, heat map of the physical layout of the flow cell comparing the efficiency of each channel, and violin plots showing base call quality against time.

Identification of species

Metagenomic sequencing data are generally associated with a high level background. In order to identify potentially pathogenic viruses and bacteria with high sensitivity and specificity, we implement a method that combines taxonomic classification, reference based mapping, and filtering.

NanoSPC first applies Centrifuge v1.0.3 (23) to classify sequencing reads to a taxonomic identifier in the centrifuge reference database (p_compressed+h+v) that comprises 20 174 complete bacterial, archaeal, viral, and human genomes corresponding to 11 539 taxonomic IDs. One primary assignment with a score >150 is reported for individual reads. Based on the centrifuge report, the pipeline selects a draft reference genome for each viral species, and maps sequencing reads to the draft reference using Minimap2 (24). In order to optimize the reference sequence for viral species, a preliminary consensus sequence for each viral species is generated using a simple majority voting method that selects the most abundant base at each genomic position. The preliminary consensus sequences are then BLASTed against a customized viral reference database to determine more optimal reference sequences. Finally, we map the reads to these references using Minimap2 for a second time. The customized viral reference database comprises of >86 000 complete genomes of viral pathogens downloaded from NIAID Virus Pathogen Database and Analysis Resource (ViPR) (25) in March 2020. We aim to update this viral reference database every four months or when significant novel pathogen species are discovered.

To distinguish species present in the data and artifacts, the pipeline implements the following filtering of the mapping results: (I) retain mapped reads with a mapping quality >50; (II) retain reads that have >80% of the bases mapped to the reference sequence (i.e. if the length of a read is 1000 bp, >800 bp are required to be mapped). Viral species with ≥2 mapped reads or one mapped read longer than 400 bp are reported.

Genome assembly and variant calling

Whole genome sequencing can provide high-resolution investigation of transmission and characterization of the spatiotemporal spread of outbreaks. While low accuracy on the sequencing read level (∼90%) is a major limitation of nanopore sequencing, genome assembly can generate high-accuracy consensus sequence with adequate sequencing depth. The full genome consensus sequences generated by NanoSPC have been used to delineate nosocomial transmission of influenza A virus and human metapneumovirus, contributing to improvement of infection prevention and control practices (26).

NanoSPC enables three variant calling modes, namely nanopolish, medaka and clair. With the nanopolish mode, NanoSPC takes both sequencing reads of the fastq format and raw signal data of the fast5 format as inputs, and applies Nanopolish (19) to detect single nucleotide variants. The pipeline considers candidate variants from the aligned reads when the variant frequency is >10% and the mapping depth is >10. To distinguish true variants and artifacts, NanoSPC keeps variants that the number of signal reads used to call the variant are >10 and the fraction of signal reads that support the variant are >75%. Finally, reads are mapped against the consensus sequence and only positions that are supported by >70% of mapped reads are kept.

While the nanopolish mode can generate consensus sequence with high accuracy, it is computationally intensive and time consuming, due to the large size of fast5 signal data. Therefore, signal based variant calling methods present a challenge to data storage, transfer, and analysis. Fortunately, with the Medaka and clair mode, NanoSPC applies Medaka (ONT) and clair (27) to detect single nucleotide variants. The advantages of these two modes is that they only take fastq format sequencing reads, and can be competitive with the nanopolish mode and being much faster.

The parameter values that we describe in the methods and functionalities section are selected based on publications (6,28) and our recent studies (26,29).

Influenza drug resistance analysis

If influenza viruses are identified from the sequencing data, the pipeline determines the type of influenza virus (A or B) and the HA and NA subtype of influenza A virus based on the reference sequences. The pipeline analyzes assembled consensus sequences of Neuraminidase and Matrix 2 genes to identify mutations that confer resistance to antiviral agents, including oseltamivir, zanamivir, and amanta-

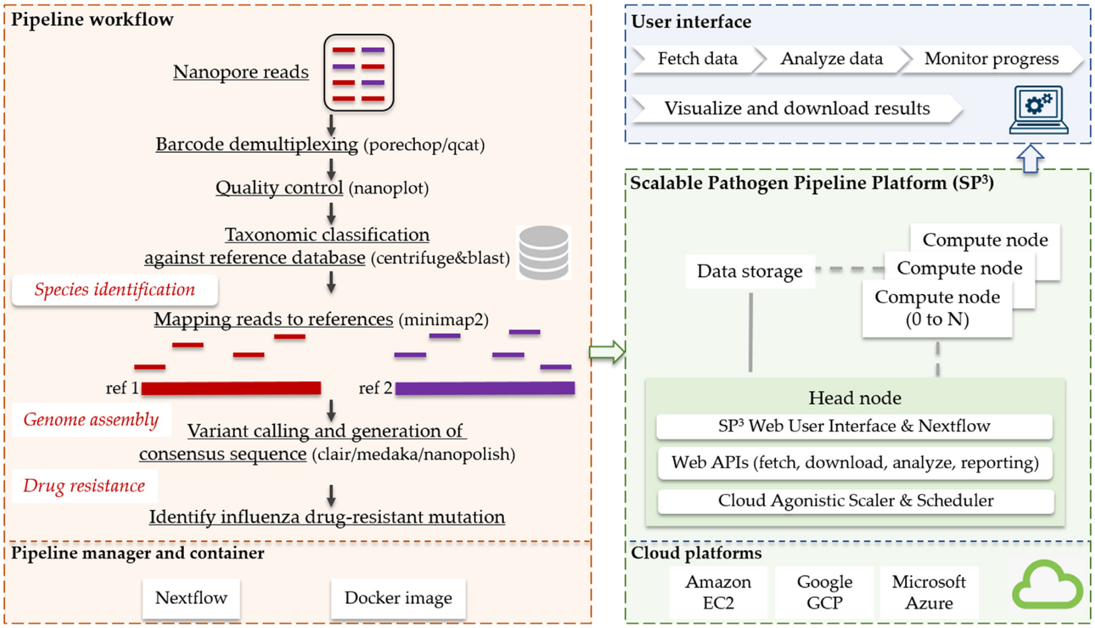

Pipeline workflow
Nanopore reads
Barcode demultiplexing (porechop/qcat)
Quality control (nanoplot)
Taxonomic classification
against reference database (centrifuge&blast)
Species identification
Mapping reads to references (minimap2)
ref 1
ref 2
Genome assembly
Variant calling and generation of
consensus sequence (clair/medaka/nanopolish)
Drug resistance
Identify influenza drug-resistant mutation
Pipeline manager and container
Nextflow
Docker image
User interface
Fetch data
Analyze data
Monitor progress
Visualize and download results
Scalable Pathogen Pipeline Platform (SP³)
Data storage
Compute node
Compute node
Compute node
(0 to N)
Head node
SP³ Web User Interface & Nextflow
Web APIs (fetch, download, analyze, reporting)
Cloud Agonistic Scaler & Scheduler
Cloud platforms
Amazon
EC2
Google
GCP
Microsoft
Azure

Figure 1. A schematic overview of NanoSPC. NanoSPC is a scalable, portable, and cloud compatible pipeline that can analyse the Nanopore metagenomic sequencing data. NanoSPC can identify potentially pathogenic viruses and bacteria simultaneously to provide comprehensive characterization of individual samples. The pipeline can also detect single nucleotide variants and assemble high quality complete consensus genome sequences. NanoSPC uses Nextflow pipeline manager and packs all the software dependencies within Docker images (red). NanoSPC can be deployed into the scalable pathogen pipeline platform, enabling elastic computing for high throughput Nanopore data on multiple cloud platforms (green). NanoSPC can be accessed via a web interface to run cloud-based analysis as well as Docker images to analyze data locally (blue).

dine. Drug resistance mutations are listed in Supplementary Table S1.

Barcode demultiplexing

The continuous improved data yield from each ONT flow cell permits users to make more efficient use and reduce cost by multiplexing several samples in a single sequencing run. Our previous work has shown that barcode contamination reads could account for <1% of the total reads from a multiplexed sequencing run (30), highlighting the need for careful barcode demultiplexing and the trade-off between sensitivity and specificity that applies to the demultiplexing methods. NanoSPC enables three demultiplexing modes, namely default-porechop, strict-porechop, and qcat. The default-porechop mode requires a barcode sequence to be present at either end of each read using porechop (https://github.com/rrwick/Porechop), thus maximizing the number of classified reads for downstream analysis. The strict-porechop mode requires the same barcode sequence to be present at both ends of each read, minimizing the number of misclassified reads. The qcat mode employs the demultiplexer offered by ONT and works similar to the default-porechop mode, due to the fact that porechop has been deprecated.

IMPLEMENTATION

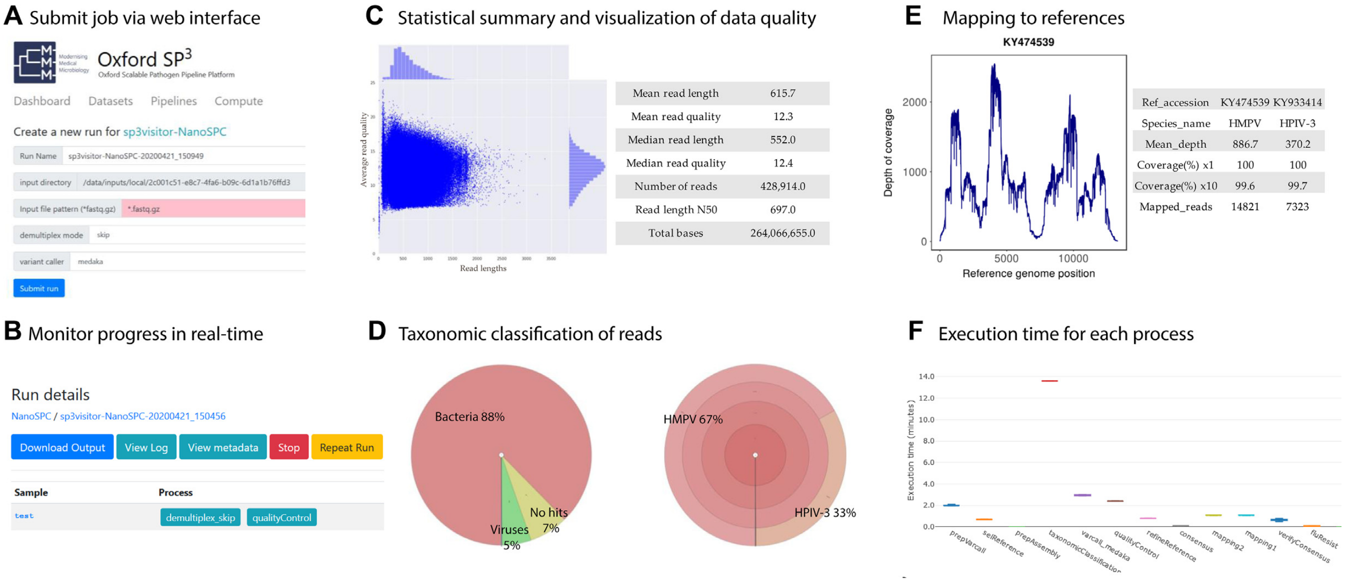
Scalable and elastic computing

In order to perform reproducible and portable analysis of sequencing data, we implement NanoSPC using Nextflow pipeline manager (31) that enables parallel processing of multiple datasets (Figure 1). Moreover, all the software dependencies of the pipeline are packed within Docker images that can be executed on a wide variety of computing infrastructures.

Scalable Pathogon Pipeline Platform (SP^3) is an open source web-based platform that we developed to host container-centric bioinformatic pipelines (32). We deploy NanoSPC to the SP^3 platform to allow the pipeline to be run at high-performance computing (HPC) clusters and major commercial cloud platforms, such as Google Cloud, Amazon Elastic Computing Cloud (Amazon EC2), Microsoft Azure and OpenStack Cloud. SP^3 has a cloud operation layer that provides interface to cloud platforms. A cloud agnostic scaler and scheduler allocate compute nodes based on CPU and memory requirement of the submitted jobs, and deallocate computational resources after the completion of the jobs. All SP^3 software are deployed to a head node equipped with Ubuntu 18.04 and Nextflow. Compute nodes are only being built and used when tasks are sent to the head node. SP^3 platform manages pipelines via a series of Web APIs, providing a range of functionalities, such as fetching data from European Nucleotide Archive (ENA) and other data sources, and excuating data analysis.

Web interface and stand-alone application

We build a web interface for users to interact with cloud platforms via SP^3 and perform data analysis (Figure 1). An authorised user can login to the SP^3 system to upload their Nanopore sequencing data and execute the pipeline. Users can monitor the progress of the submitted job in real-time. The interface displays detailed logs, running com-



Oxford SP^3
Oxford Scalable Pathogen Pipeline Platform

Dashboard Datasets Pipelines Compute

Create a new run for sp3visitor-NanoSPC

Run Name	sp3visitor-NanoSPC-20200421_150949
input directory	/data/inputs/local/2c001c51-e8c7-4fa6-b09c-6d1a1b76ffd3
Input file pattern (*fastq.gz)	*.fastq.gz
demultiplex mode	skip
variant caller	medaka

Submit run

Run details

NanoSPC / sp3visitor-NanoSPC-20200421_150456

Download Output · View Log · View metadata · Stop · Repeat Run

Sample	Process
test	demultiplex_skip qualityControl

Mean read length	615.7
Mean read quality	12.3
Median read length	552.0
Median read quality	12.4
Number of reads	428,914.0
Read length N50	697.0
Total bases	264,066,655.0

Ref_accession	KY474539	KY933414
Species_name	HMPV	HPIV-3
Mean_depth	886.7	370.2
Coverage(%) x1	100	100
Coverage(%) x10	99.6	99.7
Mapped_reads	14821	7323

Figure 2. Example showing cloud-based analysis of Nanopore metagenomic sequencing data via NanoSPC. (**A**) and (**B**) Web interfaces for executing data analysis and real-time monitoring of the progress. (**C**) Statistical summary of the data quality. (**D**) Taxonomic assignment of sequencing reads, percentage of bacterial and viral reads. (**E**) Genome coverage by mapping sequencing reads to reference sequences. (**F**) Execution time for each process in the pipeline.

mands, output files being generated, processing time, CPU usage, and memory usage for each individual process in the pipeline. Upon job completion, a complete run report displays the analysis results for each dataset. Users can choose to download the result files in bulk or only files that are of interest through running a command provided in the interface. The interface is served by the Nginx web server, configured to authenticate against web APIs, allowing user access control via web LDAP authentication. The web interface is written in python.

For users wishing to use NanoSPC to analyze data locally, we build Docker images that wrap the entire pipeline into a single environment. Users can download the Docker images and run command lines to perform the analysis.

Availability

Details of web interface for cloud-based analysis and stand-alone application are available at https://nanospc.mmmoxford.uk.

Usage

We illustrate the usage of NanoSPC with an exemplar Nanopore metagenomic sequencing dataset. The dataset contains Nanopore metagenomic sequencing reads generated from a clinical respiratory sample. The sample has been tested positive for human metapneumovirus (hmpv) in the clinical diagnostic laboratory as described in (26). Web interfaces for submitting the jobs to cloud-based analysis and real-time monitoring of the progress are shown in Figure 2A and B. Analysis of the data shows that Nanopore sequencing generated 428 914 reads with mean length of 615 bp (Figure 2C). Taxonomic classification of the sequencing reads are displayed using the Krona plots (Figure 2D). In this case, bacterial and viral reads accounted for 88% and 5% of the total sequencing reads. Among the viral reads, 67% and 33% are identified as hmpv and human parainfluenza virus type 3 (hpiv-3) reads, exemplifying simultaneous identification of multiple potentially pathogenic species from individual metagenomic sequencing dataset. Mapping to reference sequences showed that 14 821 hmpv reads cover the complete genome at mean coverage of 886, and 7323 hpiv-3 reads cover the complete genome at mean coverage of 370 (Figure 2E). The medaka mode is employed for calling variants, the total execution time is 13 mins (Figure 2F).

CONCLUSION

We report NanoSPC, a scalable, portable, and cloud compatibilable pipeline for analyzing metagenomic sequencing data generated using ONT. NanoSPC differs from other pipelines, such as NanoPipe (33), by analyzing metagenomic sequencing data and identifying a range of organisms without a priori knowledge of species contained in the data. The implementation of cloud computing enables NanoSPC to increase the ease and efficiency of analysis of high throughput Nanopore metagenomic data. The analysis results can potentially be used in multiple clinical and public health applications.

We have applied NanoSPC to identify and assemble genomes of a range of pathogen species, including influenza viruses and human metapneumovirus, from Nanopore metagenomic sequencing data of clinical samples (29). The generated sequences have been used to investigate drug resistance and genetic diversity of influenza viruses from a UK hospital during the 2018/19 influenza season. The sequences also provided high resolution characterization of nosocomial transmission of influenza A virus and human metapneumovirus, contributing to improvement of infection prevention and control practices (26).

Our future step is to include prediction of antiviral drug resistance for viral species, such as influenza A virus, and phylogenetic analysis in the pipeline. Other features including discovery of novel pathogen species, particularly viruses, are also under investigation. ONT is undergoing constant improvement in sequencing chemistry and library preparation methods. ONT has released a new sequencing chemistry (R10 version flow cell) that provides improved accuracy. Our pipeline should be able to process R10 data but validating tests have not been conducted due to data availability. We are committed to continuously update our pipeline with the development of bioinformatics tools to enhance the accuracy and efficiency of the data analysis.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

NIHR Oxford Biomedical Research Centre. Funding for open access charge: NIHR Oxford Biomedical Research Centre.

Conflict of interest statement. None declared.

REFERENCES

1. Jain,M., Koren,S., Miga,K.H., Quick,J., Rand,A.C., Sasani,T.A., Tyson,J.R., Beggs,A.D., Dilthey,A.T., Fiddes,I.T. *et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, **36**, 338–345.
2. Sakamoto,Y., Sereewattanawoot,S. and Suzuki,A. (2020) A new era of long-read sequencing for cancer genomics. *J Hum Genet*, **65**, 3–10.
3. Wouter,D.C., Arne,D.R., De Pooter,T., Svenn,D., Peter,D.R., Mojca,S., Sleegers,K. and Christine,V.B.(2019) Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Research*, **29**, 1178–1187.
4. Jain,M., Olsen,H.E., Paten,B. and Akeson,M. (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.*, **17**, 239.
5. Belser,C., Istace,B., Denis,E., Dubarry,M., Baurens,F.-C., Falentin,C., Genete,M., Berrabah,W., Chèvre,A.M., Delourme,R. *et al.* (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Na.t Plants*, **4**, 879–887.
6. Kafetzopoulou,L.E., Pullan,S.T., Lemey,P., Suchard,M.A., Ehichioya,D.U., Pahlmann,M., Thielebein,A., Hinzmann,J., Oestereich,L., Wozniak,D.M. *et al.* (2019) Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. *Science*, **363**, 74–77.
7. Quick,J., Loman,N.J., Duraffour,S., Simpson,J.T., Severi,E., Cowley,L., Bore,J.A., Koundouno,R., Dudas,G., Mikhail,A. *et al.* (2016) Real-time, portable genome sequencing for Ebola surveillance. *Nature*, **530**, 228–232.
8. Quick,J., Grubaugh,N.D., Pullan,S.T., Claro,I.M., Smith,A.D., Gangavarapu,K., Oliveira,G., Robles-Sikisaka,R., Rogers,T.F., Beutler,N.A. *et al.* (2017) Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.*, **12**, 1261–1276.
9. Chiu,C.Y. and Miller,S.A. (2019) Clinical metagenomics. *Nat. Rev. Genet.*, **20**, 341–355.
10. Schlaberg,R., Chiu,C.Y., Miller,S., Procop,G.W., Weinstock,G., Professional Practice Committee and Committee on Laboratory Practices of the American Society for Microbiology and Microbiology Resource Committee of the College of American Pathologists. (2017) Validation of metagenomic next-generation sequencing tests for universal pathogen detection. *Arch. Pathol. Lab. Med.* **141**, 776–786.
11. Wilson,M.R., Naccache,S.N., Samayoa,E., Biagtan,M., Bashir,H., Yu,G., Salamat,S.M., Somasekar,S., Federman,S., Miller,S. *et al.* (2014) Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N. Engl. J. Med.*, **370**, 2408–2417.
12. Charalampous,T., Kay,G.L., Richardson,H., Aydin,A., Baldan,R., Jeanes,C., Rae,D., Grundy,S., Turner,D.J., Wain,J. *et al.* (2019) Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat. Biotechnol.*, **37**, 783–792.
13. Greninger,A.L., Naccache,S.N., Federman,S., Yu,G., Mbala,P., Bres,V., Stryke,D., Bouquet,J., Somasekar,S., Linnen,J.M. *et al.* (2015) Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med.*, **7**, 99.
14. Kafetzopoulou,L.E., Efthymiadis,K., Lewandowski,K., Crook,A., Carter,D., Osborne,J., Aarons,E., Hewson,R., Hiscox,JA., Carroll,MW. *et al.* (2018) Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. *Euro Surveill.*, **23**, 1800228.
15. Chan,J.F.-W., Yuan,S., Kok,K.-H., To,K.K.-.W., Chu,H., Yang,J. *et al.* (2020) A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet*, **395**, 514–523.
16. Rang,F.J., Kloosterman,W.P. and de Ridder,J. (2018) From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.*, **19**, 90.
17. Li,H. (2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, **32**, 2103–2110.
18. Koren,S., Walenz,B.P., Berlin,K., Miller,J.R., Bergman,N.H. and Phillippy,A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, **27**, 722–736.
19. Loman,N.J., Quick,J. and Simpson,J.T. (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods*, **12**, 733–735.
20. Vaser,R., Sović,I., Nagarajan,N. and Šikić,M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*, **27**, 737–746.
21. Ruan,J. and Li,H. (2020) Fast and accurate long-read assembly with wtdbg2. *Nat. Methods*, **17**, 155–158.
22. De Coster,W., D'Hert,S., Schultz,D.T., Cruts,M. and Van Broeckhoven,C. (2018) NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, **34**, 2666–2669.
23. Kim,D., Song,L., Breitwieser,F.P. and Salzberg,S.L. (2016) Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.*, **26**, 1721–1729.
24. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
25. Pickett,B.E., Sadat,E.L., Zhang,Y., Noronha,J.M., Squires,R.B., Hunt,V., Liu,M., Kumar,S., Zaremba,S., Gu,Z. *et al.* (2012) ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.*, **40**, D593–D598.
26. Xu,Y., Lewandowski,K., Jeffery,K., Downs,L.O., Foster,D., Sanderson,N.D., Kavanagh,J., Vaughan,A., Salvagno,C., Vipond,R. *et al.* (2020) Nanopore metagenomic sequencing to investigate nosocomial transmission of human metapneumovirus from a unique genetic group among haematology patients in the United Kingdom. *J. Infect.*, **80**, 571–577.
27. Luo,R., Wong,C.-L., Wong,Y.-S., Tang,C.-I., Liu,C.-M., Leung,C.-M. and Lam,T.-W. (2020) Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nature Machine Intelligence* , **2**, 220–227.
28. Sanderson,N.D., Street,T.L., Foster,D., Swann,J., Atkins,B.L., Brent,A.J., McNally,M.A., Oakley,S., Taylor,A., Peto,T.E.A. *et al.* (2018) Real-time analysis of nanopore-based metagenomic sequencing from infected orthopaedic devices. *BMC Genomics*, **19**, 714.
29. Lewandowski,K., Xu,Y., Steven,T., Pullan,S.T., Lumley,S.F., Dona,F., Nicholas,S., Alison,V., Morgan,M., Bright,N. *et al.* (2019) Metagenomic nanopore sequencing of influenza virus direct from clinical respiratory samples. *J. Clin. Microbiol.* **58**, e00963-19.
30. Xu,Y., Lewandowski,K., Lumley,S., Pullan,S., Vipond,R., Carroll,M., Foster,D., Matthews,P.C., Peto,T., Crook,D. *et al.* (2018) Detection of viral pathogens with multiplex nanopore MinION sequencing: be careful with cross-talk. *Front Microbiol*, **9**, 2225.

31. Di Tommaso,P., Chatzou,M., Floden,E.W., Barja,P.P., Palumbo,E. and Notredame,C. (2017) Nextflow enables reproducible computational workflows. *Nat Biotechnol*, **35**, 316–319.
32. Yang-Turner,F., Volk,D., Fowler,P., Swann,J., Bull,M., Hoosdally,S., Connor,T., Peto,T. and Crook,D. (2019) Scalable Pathogen Pipeline Platform (SP∧3): enabling unified genomic data analysis with elastic cloud computing. *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*. 478–480.
33. Shabardina,V., Kischka,T., Manske,F., Grundmann,N., Frith,M.C., Suzuki,Y. and Makałowski,W. (2019) NanoPipe—a web server for nanopore MinION sequencing data analysis. *GigaScience*, **8**, doi:10.1093/gigascience/giy169.