Sharp *et al. BMC Res Notes (2018) 11:120*
https://doi.org/10.1186/s13104-018-3234-8

BMC Research Notes

DATA NOTE

Open Access

Oxford Screening CSF and Respiratory samples ('OSCAR'): results of a pilot study to screen clinical samples from a diagnostic microbiology laboratory for viruses using Illumina next generation sequencing

Colin Sharp[1,2†], Tanya Golubchik[3,4†], William F. Gregory[1], Anna L. McNaughton[5], Nicholas Gow[6], Mathyruban Selvaratnam[6], Alina Mirea[6], Dona Foster[7], Monique Andersson[6], Paul Klenerman[5,6], Katie Jeffery[6] and Philippa C. Matthews[5,6*]

Abstract

Objectives: There is increasing interest in the use of metagenomic (next generation sequencing, NGS) approaches for diagnosis of infection. We undertook a pilot study to screen samples submitted to a diagnostic microbiology laboratory in a UK teaching hospital using Illumina HiSeq. In the short-term, this small dataset provides insights into the virome of human respiratory and cerebrospinal fluid (CSF) samples. In the longer term, assimilating metagenomic data sets of this nature can inform optimization of laboratory and bioinformatic methods, and develop foundations for the interpretation of results in a clinical context. The project underpins a larger ongoing effort to develop NGS pipelines for diagnostic use.

Data description: Our data comprise a complete metagenomic dataset from 20 independent samples (10 CSF and 10 respiratory) submitted to the clinical microbiology laboratory for a large UK teaching hospital (Oxford University Hospitals NHS Foundation Trust). Sequences have been uploaded to the European Nucleotide Archive and are also presented as Krona plots through which the data can be interactively visualized. In the longer term, further optimization is required to better define sensitivity and specificity of this approach to clinical samples.

Keywords: Metagenomics, Next generation sequencing, Illumina, Diagnosis, Virome, Microbiome, Infection, Respiratory, CSF

Objective

Next generation sequencing (NGS) is an attractive approach to diagnosis of infection, with the potential to offer a single diagnostic pipeline to identify viruses, bacteria and fungi from a range of clinical samples [1–4]. However, there are multiple challenges in implementing such systems, and ongoing efforts are required to develop in vitro methods for handling diverse types of clinical samples, evaluate and improve sensitivity, reduce the high burden of human reads, distinguish contaminants or commensals from pathogenic organisms, and optimise positive and negative controls.

In this small pilot study, we focused on the detection of viruses from cerebrospinal fluid (CSF) and respiratory samples submitted to a routine diagnostic microbiology laboratory in order to evaluate a methods protocol and to provide a preliminary dataset for analysis, with a view

*Correspondence: philippa.matthews@ndm.ox.ac.uk
†Colin Sharp and Tanya Golubchik contributed equally to this work
[5] Nuffield Department of Medicine, Peter Medawar Building for Pathogen Research, University of Oxford, South Parks Road, Oxford OX1 3SY, UK
Full list of author information is available at the end of the article

to optimizing our laboratory approach and providing a foundation for improving bioinformatic algorithms.

A summary of this work was presented at the UK National Federation of Infection Societies (FIS) meeting, Birmingham, November 2017 [5]. We have subsequently focused specific attention on analysis of human herpes virus 6 (HHV-6) reads from within these samples, as this provides an interesting example of an organism which is widespread, potentially pathogenic [6–8] but may also be a bystander in clinical samples [9]. These results and analysis are presented in a separate manuscript [10].

Data description

Sample cohort

We randomly selected 10 cerebrospinal fluid (CSF) and 10 respiratory samples (20 different patients represented). CSF samples were submitted to the clinical diagnostic laboratory at Oxford University Hospitals NHS Foundation Trust between November 2012 and May 2014, and respiratory samples between May and December 2014. Prior to use for this research, samples had undergone routine clinical laboratory testing and were then stored at − 80 °C.

In vitro methods

A full description of laboratory methods is provided in linked data files (see 'OSCAR protocol' listed in Table 1). In brief, samples were filtered through 0.45 μm spin column filters (Merck Millipore) to remove large cellular debris and bacterial contaminants. To increase the relative amount of encapsidated viral to host nucleic acids in the sample, we pre-treated the sample with DNAse and RNAse. Nucleic acids were extracted using the QiaAmp MinElute Virus Spin Kit (Qiagen) and recovered in nuclease-free water. Reverse transcription was primed by random hexamer primers and performed using SuperScript III reagents. Sequence independent amplification of cDNA (and DNA also carried over during extraction) was carried out by an initial addition of random octamer containing primer sequences. Subsequent PCR was performed using a single primer amplification. Illumina Nextera XT libraries were made from amplified cDNAs according to the manufacturer's protocol and sequenced on the HiSeq 4000 platform with 150-base paired end reads at the Centre for Genomic Research (CGR), University of Liverpool, UK.

Bioinformatic analysis

A full description of laboratory methods is provided in linked data files (OSCAR protocol; see description and link in Table 1). In brief, the raw FASTQ files were trimmed to remove adapter sequences and to remove low quality bases. After trimming, reads < 20 base pairs were removed. The remaining reads were classified using Kraken v0.10.5-beta [11] against a reference database comprising the human genome in combination with all RefSeq genomes for viruses, bacteria and archaea. Human-tagged reads were discarded and the remainder were taken forward for analysis.

We used Kaiju [12] to confirm that the Kraken analysis was complete, using the full Genbank non-redundant protein database for viruses, bacteria and archaea. Reads were assembled de novo using metaSPAdes v3.10 [13, 14]. Assembled contigs were classified with Kraken [11], and results were visualised with Krona [15].

Limitations

This study was undertaken as a pilot exercise to underpin refinement of both laboratory methods and analysis of metagenomic data from clinical samples. On the grounds of cost, we were restricted to analysis of a small number of samples. We did not set out to derive definitive clinical diagnosis, and the data should not be used for this purpose. There are inherent difficulties with using residual clinical samples, including bias introduced into sample selection (e.g. samples from patients with a high pre-test probability of infection tend to be used up in primary clinical testing and not available for research). In archived samples, the quality of nucleic acid may deteriorate over time (this may be especially pertinent for RNA viruses).

Our methods did not include positive and negative controls. For this reason, it is difficult to assess the sensitivity with which we detected any specific virus; it is possible that the in vitro methods may have enriched or depleted particular organisms or groups of organisms. In future, positive controls can be added by spiking samples with an organism that we anticipate would not be present in human samples, or running a parallel multiplex control panel [16].

Future studies, and an accumulation of practical experience, will be required to increase the certainty with which results of NGS platforms can be interpreted. While we anticipate instances in which a specific organism can be identified from a metagenomic dataset as the cause of a clinical syndrome, there are many instances in which ambiguity may arise as a result of the difficulties in discriminating between pathogenic organisms and contaminants or bystanders. Detailed prospective studies enrolling large numbers of study subjects are the ultimate aspiration, with the aim of collecting high resolution data

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or Accession Number)	License
OSCAR protocol (methods)	File name: OSCAR protocol Details: This contains the full laboratory and bioinformatics methods Data set: Oxford Screening of CSF and Respiratory Samples ('OSCAR'): Supplementary resources for a project using next generation sequencing (NGS) for identification of viruses from clinical laboratory samples	Portable document format (.pdf)	Figshare https://dx.doi.org/10.6084/m9.figshare.5670007 (Data citation 1)	CC-BY 4.0
Benefits and challenges of meta-genomic sequencing	File name: OSCAR table benefits and challenges Details: This table summarises some of the benefits and challenges associated with the application of metagenomic approaches (next generation sequencing, NGS) to the clinical diagnosis of viral infection Data set: Oxford Screening of CSF and Respiratory Samples ('OSCAR'): Supplementary resources for a project using next generation sequencing (NGS) for identification of viruses from clinical laboratory samples	Portable document format (.pdf)	Figshare https://dx.doi.org/10.6084/m9.figshare.5670007 (Data citation 1)	CC-BY 4.0
Metadata for 20 samples used for meta-genomic sequencing	File name: OSCAR 20 samples metadata Details: This table contains the anonymised demographic and clinical details of 20 clinical samples (10 respiratory and 10 CSF), from a UK hospital microbiology laboratory, which were processed using a metagenomic pipeline Data set: Oxford Screening of CSF and Respiratory Samples ('OSCAR'): Supplementary resources for a project using next generation sequencing (NGS) for identification of viruses from clinical laboratory samples	This file is available in Microsoft excel (.xlsx) and as comma separated variables (.csv)	Figshare https://dx.doi.org/10.6084/m9.figshare.5670007 (Data citation 1)	CC-BY 4.0
Krona plots of meta-genomic data	Oxford Screening of CSF and Respiratory Samples ('OSCAR'): interactive data visualisation using Krona to display results from a pilot project using next generation sequencing (NGS) for identification of viruses from clinical laboratory samples	Hypertext Markup Language (.html)	Figshare https://dx.doi.org/10.6084/m9.figshare.5712091 (Data citation 2)	CC-BY 4.0

Table 1 (continued)

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or Accession Number)	License
Sequence data	HHV6 reads and metagenomic data from clinical samples	Metagenomic sequences: Fastq HHV-6 reads are available in BAM and Fastq format	European Molecular Biology Laboratory (EMBL); European Nucleotide Archive (primary Accession PRJEB22949) (Data citation 3). Individual accession numbers (Data citations 4–27): SAMEA104355484; SAMEA104355485; SAMEA104355486 ; SAMEA104355487; SAMEA104421606; SAMEA104421607; SAMEA104421608; SAMEA104421609; SAMEA104421610; SAMEA104421611; SAMEA104355484; SAMEA104421613; SAMEA104421614; SAMEA104421615; SAMEA104421616; SAMEA104421617 ; SAMEA104421618; SAMEA104421619; SAMEA104421620; SAMEA104421621; SAMEA104355485; SAMEA104355486; SAMEA104355487; SAMEA104421625	Open access in accordance with ENA policy (https://www.ebi.ac.uk/ena/standards-and-policies)

that include medical history, other laboratory results, imaging, treatment and follow-up data.

Abbreviations

CSF: cerebrospinal fluid; DNA: deoxyribonucleic acid; ENA: European Nucleotide Archive; NGS: next generation sequencing; RNA: ribonucleic acid.

Authors' contributions

CS and PM designed the experiments and applied for funding and ethics approval with support from PK. MS, AM and NG collected laboratory samples and clinical data with support from KJ and MA. CS and WG undertook experimental lab work. CS, TG, ALM and PM analyzed and interpreted the data. PM wrote the manuscript with expert input from ALM, TG, MA, DF, KJ, and PK. All authors read and approved the final manuscript.

Author details

[1] The Roslin Institute, University of Edinburgh, Easter Bush, Midlothian, Edinburgh EH25 9RG, Scotland, UK. [2] Edinburgh Genomics, Ashworth Laboratories, University of Edinburgh, Edinburgh EH9 3FL, Scotland, UK. [3] The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [4] Big Data Institute, University of Oxford, Old Road, Oxford OX3 7FZ, UK. [5] Nuffield Department of Medicine, Peter Medawar Building for Pathogen Research, University of Oxford, South Parks Road, Oxford OX1 3SY, UK. [6] Department of Infectious Diseases and Microbiology, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Headley Way, Headington, Oxford OX3 9DU, UK. [7] NIHR Biomedical Research Centre, University of Oxford, John Radcliffe Hospital, Headley Way, Headington, Oxford OX3 9DU, UK.

Acknowledgements

We are grateful to the Centre for Genomic Research (CGR), University of Liverpool, UK for running the HiSeq platform. Ongoing work to develop and improve metagenomic platforms for clinical laboratory diagnosis is being undertaken with the Modernising Medical Microbiology Research Group funded by the NIHR BRC.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The data described in this Data note can be freely and openly accessed on Figshare (https://doi.org/10.6084/m9.figshare.5670007 and https://doi.org/10.6084/m9.figshare.5712091) and on the European Molecular Biology Laboratory (EMBL) European Nucleotide Archive (https://www.ebi.ac.uk/ena/data/search?query=PRJEB22949).

Consent for publication

Not applicable.

Data citations

1. Sharp C, Golubchik T, Gregory WF, McNaughton A, Gow N, Mirea A, Selvaratnam M, Foster D, Andersson M, Klenerman P, Jeffery K, Matthews P. Oxford Screening of CSF and Respiratory Samples ('OSCAR'): Supporting data files for a project using Next Generation Sequencing (NGS) for identification of viruses from clinical laboratory samples. https://dx.doi.org/10.6084/m9.figshare.5670007.

2. Golubchik T, Sharp C, Gregory WF, McNaughton A, Gow N, Mirea A, Selvaratnam M, Foster D, Andersson M, Jeffery K et al. Oxford Screening of CSF and Respiratory Samples ('OSCAR'): interactive data visualisation using Krona to display results from a pilot project using Next Generation Sequencing (NGS) for identification of viruses from clinical laboratory samples. https://dx.doi.org/10.6084/m9.figshare.5712091.

3. European Nucleotide Archive; PRJEB22949 (2017).

4. ENA Sequence Read Archive SAMEA104355484 (2017).

5. ENA Sequence Read Archive SAMEA104355485 (2017).

6. ENA Sequence Read Archive SAMEA104355486 (2017).

7. ENA Sequence Read Archive SAMEA104355487 (2017).

8. ENA Sequence Read Archive SAMEA104421606 (2017).

9. ENA Sequence Read Archive SAMEA104421607 (2017).

10. ENA Sequence Read Archive SAMEA104421608 (2017).

11. ENA Sequence Read Archive SAMEA104421609 (2017).

12. ENA Sequence Read Archive SAMEA104421610 (2017).

13. ENA Sequence Read Archive SAMEA104421611 (2017).

14. ENA Sequence Read Archive SAMEA104355484 (2017).

15. ENA Sequence Read Archive SAMEA104421613 (2017).

16. ENA Sequence Read Archive SAMEA104421614 (2017).

17. ENA Sequence Read Archive SAMEA104421615 (2017).

18. ENA Sequence Read Archive SAMEA104421616 (2017).

19. ENA Sequence Read Archive SAMEA104421617 (2017).

20. ENA Sequence Read Archive SAMEA104421618 (2017).

21. ENA Sequence Read Archive SAMEA104421619 (2017).

22. ENA Sequence Read Archive SAMEA104421620 (2017).

23. ENA Sequence Read Archive SAMEA104421621 (2017).

24. ENA Sequence Read Archive SAMEA104355485 (2017).

25. ENA Sequence Read Archive SAMEA104355486 (2017).

26. ENA Sequence Read Archive SAMEA104355487 (2017).

27. ENA Sequence Read Archive SAMEA104421625 (2017).

Ethics approval and consent to participate

This study was approved by a UK Research Ethics Committee (REC Reference 14/LO/1077) in July 2014. We did not seek informed consent from the patients, as we collected no identifying patient data, the results were obtained retrospectively (so were not relevant in informing clinical decision-making), and the focus of the study was on methods development.

Funding

The experimental work was supported by a small project grant from the British Infection Association (Awarded to PM in 2014). PM has subsequently been funded by the Wellcome Trust during the analysis phase of this project (Grant Number 110110). DF is supported by the NIHR Biomedical Research Centre (BRC). The funding bodies had no role in the design of the study and collection, analysis, interpretation of data or writing the manuscript.

Publisher's Note

Received: 20 December 2017 Accepted: 6 February 2018
Published online: 09 February 2018

References

1. Street TL, Sanderson ND, Atkins BL, Brent AJ, Cole K, Foster D, McNally MA, Oakley S, Peto L, Taylor A, et al. Molecular diagnosis of orthopedic-device-related infection directly from sonication fluid by metagenomic sequencing. J Clin Microbiol. 2017;55(8):2334–47.
2. Graf EH, Simmon KE, Tardif KD, Hymas W, Flygare S, Eilbeck K, Yandell M, Schlaberg R. Unbiased detection of respiratory viruses by use of RNA sequencing-based metagenomics: a systematic comparison to a commercial PCR panel. J Clin Microbiol. 2016;54(4):1000–7.
3. Greninger AL, Zerr DM, Qin X, Adler AL, Sampoleo R, Kuypers JM, Englund JA, Jerome KR. Rapid metagenomic next-generation sequencing during an investigation of hospital-acquired human parainfluenza virus 3 infections. J Clin Microbiol. 2017;55(1):177–82.
4. Perlejewski K, Popiel M, Laskus T, Nakamura S, Motooka D, Stokowy T, Lipowski D, Pollak A, Lechowicz U, Cortes KC, et al. Next-generation sequencing (NGS) in the identification of encephalitis-causing viruses: unexpected detection of human herpesvirus 1 while searching for RNA pathogens. J Virol Methods. 2015;226:1–6.
5. McNaughton A, Golubchik T, Sharp C, Gregory WF, Selvaratnam M, Mirea A, Klenerman P, Jeffery K, Matthews PC. Oxford screening of CSF and Respiratory Samples ('OSCAR'): comparison of routine laboratory diagnostics, multiplex PCR and next generation sequencing for identification of viruses from clinical samples. F1000Research. 2017;6:2130 (poster). https://doi.org/10.7490/f1000research.1115150.1.
6. Colombier MA, Amorim S, Salmona M, Thieblemont C, Legoff J, Lafaurie M. HHV-6 reactivation as a cause of fever in autologous hematopoietic stem cell transplant recipients. J Infect. 2017;75(2):155–9.

7. de Pagter PJ, Schuurman R, Keukens L, Schutten M, Cornelissen JJ, van Baarle D, Fries E, Sanders EA, Minnema MC, van der Holt BR, et al. Human herpes virus 6 reactivation: important predictor for poor outcome after myeloablative, but not non-myeloablative allo-SCT. Bone Marrow Transplant. 2013;48(11):1460–4.
8. Inazawa N, Hori T, Yamamoto M, Hatakeyama N, Yoto Y, Nojima M, Yasui H, Suzuki N, Shimizu N, Tsutsumi H. HHV-6 encephalitis may complicate the early phase after allogeneic hematopoietic stem cell transplantation: detection by qualitative multiplex PCR and subsequent quantitative real-time PCR. J Med Virol. 2016;88(2):319–23.
9. Pellett PE, Ablashi DV, Ambros PF, Agut H, Caserta MT, Descamps V, Flamand L, Gautheret-Dejean A, Hall CB, Kamble RT, et al. Chromosomally integrated human herpesvirus 6: questions and answers. Rev Med Virol. 2012;22(3):144–55.
10. Sharp C, Golubchik T, Gregory WF, McNaughton A, Gow N, Selvaratnam M, Mirea A, Foster D, Andersson M, Klenerman P et al. Human Herpes Virus 6 (HHV-6)-pathogen or passenger? A pilot study of clinical laboratory data and next generation sequencing. bioRxiv. 2017. https://doi.org/10.1101/236083.
11. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46.
12. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016;7:11257.
13. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
14. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27(5):824–34.
15. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinform. 2011;12:385.
16. Mee ET, Preston MD, Minor PD, Schepelmann S, Participants CSS. Development of a candidate reference material for adventitious virus detection in vaccine and biologicals manufacturing by deep sequencing. Vaccine. 2016;34(17):2035–43.