Oxford Screening of CSF And Respiratory Samples ('OSCAR'): Pilot protocol for next generation sequencing (NGS) for identification of viruses from clinical laboratory samples

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

[*] These two authors contributed equally to this work

[1] Roslin Institute, University of Edinburgh, Edinburgh EH25 9RG, UK
[2] Edinburgh Genomics, University of Edinburgh, Edinburgh EH9 3FL, UK
[3] Big Data Institute, University of Oxford, Old Road Campus, Oxford OX3 7FZ, UK
[4] Nuffield Department of Medicine, Peter Medawar Building, South Parks Road, Oxford OX1 3SY UK
[5] Department of Infectious Diseases and Microbiology, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Headley Way, Headington, Oxford OX3 9DU, UK
[6] NIHR British Research Centre, Nuffield Department of Medicine, John Radcliffe Hospital, Headley Way, Headington, Oxford OX3 9DU, UK

[#] Corresponding author: philippa.matthews@ndm.ox.ac.uk

BACKGROUND

In an era of rapidly evolving techniques for molecular diagnosis of pathogens, careful collaborative efforts are required to develop robust pipelines for clinical laboratory approaches and data analysis. We here describe the protocols in a small pilot study to apply next generation sequencing (NGS) to anonymised samples collected in a UK hospital, testing cerebrospinal fluid (CSF) and respiratory samples (n=10 of each). This work provides foundations on which diagnostic metagenomic platforms can be improved and developed. There is increasing interest and investment in developing next generation sequencing (NGS) for diagnostic testing in the clinical microbiology arena (1-3), but a number of challenges mean that additional work is required before such approaches can be applied in clinical diagnostics. The benefits and challenges are summarized in Suppl. Table 1 (see associated file set).

AIMS

This project was designed as a pilot study centred in a UK teaching hospital microbiology laboratory, to screen CSF and respiratory samples using Illumina NGS. We set out to inform methodological improvements in laboratory and bioinformatic approaches to viral diagnostics.

METHODS

Ethics

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.

Sample and data collection

Samples were obtained from the CQC accredited diagnostic microbiology laboratory at Oxford University Hospitals NHS Foundation Trust (http://www.ouh.nhs.uk/), a large teaching hospital accepting secondary and tertiary referrals in the South-East of England. We selected at random 10 cerebrospinal fluid (CSF) samples and 10 respiratory samples (from a cohort comprising nose/throat swabs (n=24), nasopharyngeal aspirates, NPA (n=45), and broncho-alveolar lavage samples, BAL (n=31)) that had completed routine diagnostic testing (metadata for these 20 samples is available in Suppl Table 2). We accessed the samples at this point prior to discard

from the clinical laboratory, at which point they had been stored at -80°C for a period of 3-6 months. CSF samples had been submitted to the laboratory between November 2012 and May 2014, and respiratory samples between May and December 2014. Twenty independent patients were represented.

We collected data from the hospital electronic patient record (EPR), recording anonymised data including patient age (in years), location within the hospital, and clinical information provided by the requester at the time of submitting the sample to the microbiology laboratory. Results obtained by diagnostic testing in the clinical microbiology laboratory were recorded for each sample, and for CSF samples we also recorded the differential white cell count.

CSF and respiratory samples

Prior to use for this research, samples had undergone clinical laboratory testing as follows:

i. ***CSF samples:*** total and differential cell counts were ascertained by light microscopy, followed by bacterial culture. PCR for viral causes of meningitis/encephalitis was also undertaken (multiplex for Herpes Simplex Virus (HSV), Varicella Zoster Virus (VZV), Mumps, Enterovirus and Parechovirus); (QIAsymphony DSP Virus/Pathogen Mini kit, version1, FTD® Viral meningitis kit).

ii. ***Respiratory samples:*** following bacterial culture and PCR for Influenza A and B and Respiratory Syncytial Virus (RSV), an extended PCR panel was undertaken (multiplex for Influenza A, Influenza B, Rhinovirus, Coronaviruses, Parainfluenza 1-4, Adenovirus, Metapneumovirus, RSV, Enterovirus, Parechovirus, Bocavirus and *Mycoplasma pneumoniae*); (QIAsymphony DSP Virus/Pathogen Mini kit, version1, FTD® Respiratory pathogens 21).

The only exclusion criterion was inadequate sample volume (<250 µl). For each sample, we documented patient age group and clinical location, and recorded the clinical information that had been supplied with the sample and microbiology laboratory data generated as part of the primary clinical testing pipeline.

Additional experimental testing was undertaken at the Roslin Institute, University of Edinburgh. 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, 15 µL TURBO DNA-free DNase (Life Technologies), 25-60 µL of 10x TURBO DNAase buffer and 15 µL RNase A (Thermo Fisher Scientific) was added to 250-600 µL of filtered sample and the reaction was incubated for 90 min at 37ºC. Nucleic acids were extracted from the full volume of 305-690 µl sample using the QiaAmp MinElute Virus Spin Kit (Qiagen) and recovered in 30 µl of nuclease-free water. Reverse transcription was primed by random hexamer primers and performed using SuperScript III reagents (Invitrogen) according to manufacturer's instructions. As the RNA yields in the filtered and nuclease treated samples were too low for accurate quantification, a volume of 22 ul of extracted RNA was added to each 40 ul SuperScript III reaction. 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 (RP1: GCC GGA GCT CTG CAG ATA TCN NNN NNN N) by two rounds of Klenow 3'-5' exo$^-$ (New England Biolabs) polymerisation according to manufacturer's instructions.

Eight subsequent replicate PCRs were performed on each of the samples to obtain a maximal output from each sample. The reactions used a single primer amplification (RP2: GCC GGA GCT CTG CAG ATA TC) with AmpliTaq Gold DNA Polymerase LD under the following cycling conditions: initial denaturation at 95°C for 5 minutes followed by 5 cycles of [60 seconds at 95°C, 60 seconds at 55°C and 2 minutes at 72°C] and 35 cycles of [30 seconds at 95°C, 30 seconds at 55°C and 2 minutes at 72°C]. Illumina Nextera XT libraries were made from amplified cDNAs according to the manufacturer's protocol (1ng nucleic acid) 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

The raw FASTQ files were trimmed to remove Illumina adapter sequences and custom adapter sequence "GCCGGAGCTCTGCAGATATC" using Cutadapt version 1.2.1 (4). The option "-O 3" was set, so the 3' end of any reads which matched the adapter sequence over at least 3 bp was trimmed off. The reads were further trimmed to remove low quality bases, using Sickle version 1.2 (5) with a minimum window quality score of 20. After trimming, reads shorter than 20bp were removed. This approach to cleaning the sequence data is in keeping with that described by other authors for metagenomic data (6).

Trimmed, quality-filtered reads were classified using Kraken version 0.10.5-beta (7) 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 (8) to confirm that the Kraken analysis was complete, using the full Genbank non-redundant protein database for viruses, bacteria and archaea. Additionally, reads were assembled *de novo* using metaSPAdes v3.10 (9, 10). Assembled contigs were classified with Kraken (7), and all Kraken results were visualised with Krona (11).

For estimates of genomic coverage of HHV-6, reads were mapped with BWA-MEM (12) to the majority-rule consensus of reference genomes for HHVA (GenBank IDs KP257584.1, KC465951.1 and KJ123690.1) and HHVB (Genbank ID AF157706 and AB021506). Reads mapped with a minimum quality of 30 (q30) were extracted with samtools v1.3 (13) and deduplicated using MarkDuplicates [Picard tools v2.0.1 [http://broadinstitute.github.io/picard/].

RESULTS

Results of Illumina sequencing from our 20 samples can be accessed in two formats:

i. html file displaying data as interactive Krona plots (OSCAR Suppl Fig 1 in this file set);
ii. Uploaded data available at European Nucleotide Archive (https://www.ebi.ac.uk/ena); European Molecular Biology Laboratory (EMBL); European Nucleotide Archive (primary accession PRJEB22949) (14).

DISCUSSION

Other recent publications have also demonstrated the potential utility of applying NGS technology to respiratory (15-19) and CSF (20-22) samples. This is a small exploratory study that sets out to inform a better understanding of methods for laboratory diagnostics of viral infection, with a focus on metagenomic pipelines.

FUNDING

PCM has received salary support from NIHR and the Wellcome Trust (Grant ref. 110110/Z/15/Z) during this project. Consumables and clinical biomedical scientist time was funded by a project grant from the British Infection Association to PCM in 2014.

REFERENCES

1. Mulcahy-O'Grady H, Workentine ML. The Challenge and Potential of Metagenomics in the Clinic. Front Immunol. 2016;7:29.

2. Quinones-Mateu ME, Avila S, Reyes-Teran G, Martinez MA. Deep sequencing: becoming a critical tool in clinical virology. J Clin Virol. 2014;61(1):9-19.

3. Street TL, Sanderson ND, Atkins BL, Brent AJ, Cole K, Foster D, et al. Molecular diagnosis of orthopaedic device infection direct from sonication fluid by metagenomic sequencing. bioRxiv.doi: https://doi.org/10.1101/118026.

4. Martin M. Cutadapt v. 1.2.1. Trim adapters from high-throughput sequencing reads. http://code.google.com/p/cutadapt/ 2010-2012.

5. Joshi N, Fass J. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files [Software]. Available at https://github.com/najoshi/sickle. 2011.

6. Arkhipova K, Skvortsov T, Quinn JP, McGrath JW, Allen CC, Dutilh BE, et al. Temporal dynamics of uncultured viruses: a new dimension in viral diversity. ISME J. 2018;12(1):199-211.

7. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46.

8. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016;7:11257.

9. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-77.

10. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27(5):824-34.

11. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011;12:385.

12. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.

13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

14. European Nucleotide Archive; https://www.ebi.ac.uk/ena 2017 [Available from: https://www.ebi.ac.uk/ena.

15. Thorburn F, Bennett S, Modha S, Murdoch D, Gunson R, Murcia PR. The use of next generation sequencing in the diagnosis and typing of respiratory infections. J Clin Virol. 2015;69:96-100.

16. Zou X, Tang G, Zhao X, Huang Y, Chen T, Lei M, et al. Simultaneous virus identification and characterization of severe unexplained pneumonia cases using a metagenomics sequencing technique. Sci China Life Sci. 2017;60(3):279-86.

17. Graf EH, Simmon KE, Tardif KD, Hymas W, Flygare S, Eilbeck K, et al. 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.

18. Lewandowska DW, Zagordi O, Zbinden A, Schuurmans MM, Schreiber P, Geissberger FD, et al. Unbiased metagenomic sequencing complements specific routine diagnostic methods and increases chances to detect rare viral strains. Diagn Microbiol Infect Dis. 2015;83(2):133-8.

19. Nakamura S, Yang CS, Sakon N, Ueda M, Tougan T, Yamashita A, et al. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. PLoS One. 2009;4(1):e4219.

20. Bukowska-Osko I, Perlejewski K, Nakamura S, Motooka D, Stokowy T, Kosinska J, et al. Sensitivity of Next-Generation Sequencing Metagenomic Analysis for Detection of RNA and DNA Viruses in Cerebrospinal Fluid: The Confounding Effect of Background Contamination. Adv Exp Med Biol. 2017;944:53-62.

21. Perlejewski K, Popiel M, Laskus T, Nakamura S, Motooka D, Stokowy T, 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.

22. Phan TG, Mori D, Deng X, Rajindrajith S, Ranawaka U, Fan Ng TF, et al. Small circular single stranded DNA viral genomes in unexplained cases of human encephalitis, diarrhea, and in untreated sewage. Virology. 2015;482:98-104.