www.nature.com/scientificreports

SCIENTIFIC REPORTS

natureresearch

Check for updates

OPEN

A metagenomics-based diagnostic approach for central nervous system infections in hospital acute care setting

Mohammad Rubayet Hasan[1,2,3], Sathyavathi Sundararaju[3], Patrick Tang[2,3], Kin-Ming Tsui[2,3], Andres Perez Lopez[2,3], Mohammad Janahi[2,3], Rusung Tan[2,3] & Peter Tilley[4,5]

The etiology of central nervous system (CNS) infections such as meningitis and encephalitis remains unknown in a large proportion of cases partly because the diversity of pathogens that may cause CNS infections greatly outnumber available test methods. We developed a metagenomic next generation sequencing (mNGS)-based approach for broad-range detection of pathogens associated with CNS infections suitable for application in the acute care hospital setting. The analytical sensitivity of mNGS performed on an Illumina MiSeq was assessed using simulated cerebrospinal fluid (CSF) specimens (n = 9). mNGS data were then used as a training dataset to optimize a bioinformatics workflow based on the IDseq pipeline. For clinical validation, residual CSF specimens (n = 74) from patients with suspected CNS infections previously tested by culture and/or PCR, were analyzed by mNGS. In simulated specimens, the NGS reads aligned to pathogen genomes in IDseq were correlated to qPCR C_T values for the respective pathogens ($R = 0.96$; $p < 0.0001$), and the results were highly specific for the spiked pathogens. In clinical samples, the diagnostic accuracy, sensitivity and specificity of the mNGS with reference to conventional methods were 100%, 95% and 96%, respectively. The clinical application of mNGS holds promise to benefit patients with CNS infections of unknown etiology.

Central nervous system (CNS) infections such as meningitis and encephalitis are potentially life threatening diseases caused by a myriad of infectious pathogens. Besides high rates of mortality, meningitis and encephalitis are major causes of morbidity, and permanent disabilities such as brain damage, hearing loss, and learning disabilities can result from CNS infections[1–5]. A specific etiologic agent cannot be identified in 15–60% of cases of meningitis and up to 70% of encephalitis[6, 7]. Clinical management of meningitis and encephalitis is highly dependent on early and rapid detection of underlying causes of the disease, so that appropriate antimicrobial or anti-viral therapy can be instituted in a timely manner. Specific diagnosis is also important to avoid unnecessary treatment and hospitalization of patients with self-limiting forms of viral meningitis to minimize potential harm and unnecessary cost to patients[6].

Current diagnostic test methods for CNS infections include CSF Gram staining, CSF cell count, glucose, and protein measurements and biomarkers such as procalcitonin (PCT) and lactate. These tests are generally performed to distinguish between bacterial versus viral infections, and they are not specific for any causative pathogens. Bacteriological culture or PCR testing to detect specific pathogens in cerebrospinal fluid (CSF) are currently the most important methods for the diagnosis of CNS infections. However, a large number pathogens known to cause meningitis and encephalitis cannot be routinely cultured, and most molecular tests are targeted to common pathogens only. A broad-range, unbiased method to identify all pathogens in CSF would markedly improve the management of patients who are critically ill with undiagnosed disease.

Advances in genomic approaches—particularly in sequencing technologies—are being applied in many research and clinical settings. For example, next generation sequencing (NGS) technology, which is capable of deciphering millions of DNA and RNA sequences in parallel, has shown promise for detecting pathogens in clinical samples[8]. In a recent online survey, with infectious diseases physicians, microbiologists and other associated

[1]Department of Pathology, Sidra Medicine, Level 2M, Office H2M-24093, PO BOX 26999, Doha, Qatar. [2]Weill Cornell Medical College in Qatar, Doha, Qatar. [3]Sidra Medicine, Doha, Qatar. [4]British Columbia Children's Hospital, Vancouver, BC, Canada. [5]University of British Columbia, Vancouver, BC, Canada. ✉email: mhasan@sidra.org

professionals as participants, all respondents predict that NGS will find some use in the clinical microbiology laboratories within the next 5–10 years[9]. Application of NGS in clinical microbiology laboratories include whole genome sequencing (WGS) of purified bacterial isolates for identification, typing, detection of antibiotic resistance, virulence profiling and epidemiological surveillance for infection control[9]. On the other hand, NGS based metagenomic sequencing, which involves sequencing of all DNA content in the sample, has been applied mostly in research settings for microbiome studies as well as pathogen detection or characterization directly from clinical specimens[10–17]. However, application of this technology in acute care diagnostic microbiology is limited due to its higher cost compared to conventional microbiological methods, lack of standardized methods, data interpretation challenges and the burden of analyzing and storing large datasets of sequences.

Recently, a heightened interest in the application of NGS in clinical microbiology laboratories has led to an increased number of studies aimed to optimize, standardize and validate mNGS for the diagnosis of CNS infections[18–23]. The results of some of these studies are highly encouraging, particularly because of the ability of NGS to detect pathogens that are unidentifiable by conventional testing and their potential clinical impact in directing appropriate antimicrobial therapy. However, these methods are difficult to implement in acute care hospital settings because of complex laboratory workflows, too few specimens for batching, and lack of bioinformatics expertise for data analysis. We hypothesized that a simplified, low throughput approach adapted for implementation in an acute care diagnostic microbiology laboratory will provide actionable, clinically useful data with faster turnaround time and thus benefit a larger patient population with CNS infections of unknown etiology.

To this end, the benchtop sequencing platform, Illumina MiSeq, is a relatively lower cost instrument compared to larger Illumina systems, that is suitable for low-throughput applications such as clinical metagenomic sequencing of single samples or applications such as small genome sequencing, targeted gene sequencing and 16S rRNA sequencing. MiSeqDx version of the instrument is also the first FDA-regulated, CE-IVD-marked, NGS platform for in vitro diagnostic (IVD) testing. MiSeq system offers simpler NGS library preparation protocol, very low input DNA requirement, high quality data and faster turnaround time. This platform is comparable to larger Illumina platform in its ability to sequence microbes from host-associated and environmental specimens[24]. However, use of the MiSeq platform has not been standardized for metagenomic pathogen detection in clinical samples. In this study, we have determined that the analytical sensitivity of a MiSeq-based, shotgun metagenomic sequencing approach to detect bacterial and viral pathogens in cerebrospinal fluid (CSF) is comparable to PCR assays and then optimized a bioinformatic approach for high specificity based on a publicly accessible software platform for metagenomic detection of pathogens (https://idseq.net)[25, 26]. The optimized approach was then applied to a set of retrospectively collected and previously tested CSF specimens ($n = 74$) with an aim to describe the diagnostic accuracy, sensitivity and specificity of mNGS approach with reference to the standard methods. Our results suggest that mNGS-based testing can be implemented in clinical microbiology laboratories for the routine diagnosis of CNS infections, when broad-range detection of pathogens is required.

Results

Establishment of mNGS laboratory workflow.

With an aim to establish a Illumina MiSeq based workflow for the detection of CNS pathogens, and to assess the analytical sensitivity of our mNGS approach, a set of CSF specimens (Training set) that were negative for CNS pathogens by routine laboratory investigations were spiked with a range of reference bacterial or viral strains at varying concentrations and assessed by both qPCR and mNGS. Negative CSF specimens (not spiked) and nuclease free water (NFW) were also assessed by mNGS to determine the level of background noise and contamination. Spiking was done keeping in mind that: (i) the approximate titers of pathogens range from very low to medium to very high concentrations; (ii) spiked organisms are typical CNS pathogens; and (iii) spiked organisms represent various groups of bacteria and viruses such as gram positive and gram negative bacteria and enveloped and non-enveloped viruses. In order to reduce the cost of sequencing, multiple pathogens were spiked into the same specimen (Table 1). DNA concentrations in the CSF specimens, as well as the resulting NGS libraries prepared using the Nextera XT kit, were highly variable, but sufficient for sequencing (Supplementary Table 1). Also, the nucleotide (NT) reads associated with the internal control plasmid spiked into the specimens prior to extraction was highly variable with an average of about 3,070 reads per sample, ranging from 55 to 21,640 reads. However, the NGS reads aligned to pathogen genomes (NT reads) in IDseq were highly comparable to qPCR C_T (cycle threshold) values for the respective pathogens. While a statistically significant correlation ($R = 0.63$; $p < 0.001$) was observed between approximate titers of spiked pathogens with their respective NT reads, the qPCR C_T values for different pathogens were more precisely correlated to their mNGS NT reads ($R = 0.96$; $p < 0.0001$) (Fig. 1A and B). mNGS assay detected all spiked pathogens that were detectable by qPCR including a bacterium that was spiked at 100 CFU/ml final concentration (Table 1). The limit of detection of mNGS assay at PCR C_T 40 was estimated to be 0.93 $\log_{10}$(CFU/ml) (Supplementary Fig. 1). The $\log_{10}$(NT reads) for different pathogens by mNGS assay was reproducible with %CV ranging from 2.2 to 11% (Table 2). Based on these results, a simplified laboratory workflow was set for clinical validation (Fig. 2).

Establishment of mNGS bioinformatics workflow.

While the DNA associated with all spiked pathogens were detected by mNGS using the IDseq software with high sensitivity, the specificity of mNGS results was poor using the default settings. In this setting, sequence reads from specimens mapped to a large number of taxa that include both common contaminants as well as bioinformatic artifacts originating from poor quality alignments making it difficult to differentiate true positive results from false positive results (Supplementary Fig. 2). Along with the spiked specimens, this is also reflected in the IDseq results of known negative CSF samples as well as in the negative control water sample. Compared to the IDseq results, Metaphlan2 results were relatively more

Specimen no	Organisms spiked	Approximate titer Log(CFU or TCID$_{50}$)/ml	qPCR result (C$_T$)	IDseq NT reads
Spike 1	Human adenovirus type 7	4.7 TCID$_{50}$/ml	21.0	198,994
	Herpes simplex virus 2	3.7 TCID$_{50}$/ml	24.6	10,842
	Haemophilus influenzae	5.9 CFU/ml	27.7	83,901
	Escherichia coli	5.9 CFU/ml	29.2	21,971
	Streptococcus pneumoniae	5.9 CFU/ml	31.9	194
Spike 2	Herpes simplex virus 2	3.5 TCID$_{50}$/ml	22.3	121,258
	Neisseria meningitidis	6.6 CFU/ml	23.7	800,771
	Streptococcus agalactiae	6.0 CFU/ml	30.2	20,110
	Escherichia coli	4.0 CFU/ml	34.6	2,643
	Haemophilus influenzae	2.0 CFU/ml	Undetermined	48
Spike 3	Herpes simplex virus 2	3.5 TCID$_{50}$/ml	23.3	188,394
	Neisseria meningitidis	6.6 CFU/ml	24.2	1,744,667
	Streptococcus agalactiae	6.0 CFU/ml	32.6	12,761
	Escherichia coli	4.0 CFU/ml	34.3	6,825
	Haemophilus influenzae	2.0 CFU/ml	Undetermined	147
Spike 4	Herpes simplex virus 2	3.5 TCID$_{50}$/ml	23.2	141,614
	Neisseria meningitides	6.6 CFU/ml	24.3	1,040,079
	Streptococcus agalactiae	6.0 CFU/ml	31.3	19,255
	Escherichia coli	4.0 CFU/ml	35	3,276
	Haemophilus influenzae	2.0 CFU/ml	Undetermined	88
Spike 5	Herpes simplex virus 2	3.5 TCID$_{50}$/ml	24.3	217,550
	Neisseria meningitides	6.6 CFU/ml	25.1	1,720,187
	Streptococcus agalactiae	6.0 CFU/ml	33.9	5,787
	Escherichia coli	4.0 CFU/ml	35.4	6,825
	Haemophilus influenzae	2.0 CFU/ml	Undetermined	130
NEG1	None	0	–	–
NEG2	None	0	–	–
NEG3	None	0	–	–
NEG4	None	0	–	–
NFW	None	0	–	–

Table 1. Composition, approximate titer, qPCR C$_T$ values and mNGS sequence read outputs of spiked organisms in the simulated CSF specimens. *NFW* nuclease free water, *NT* nucleotide, *TCID* tissue culture infectious dose, *CFU* colony forming unit.

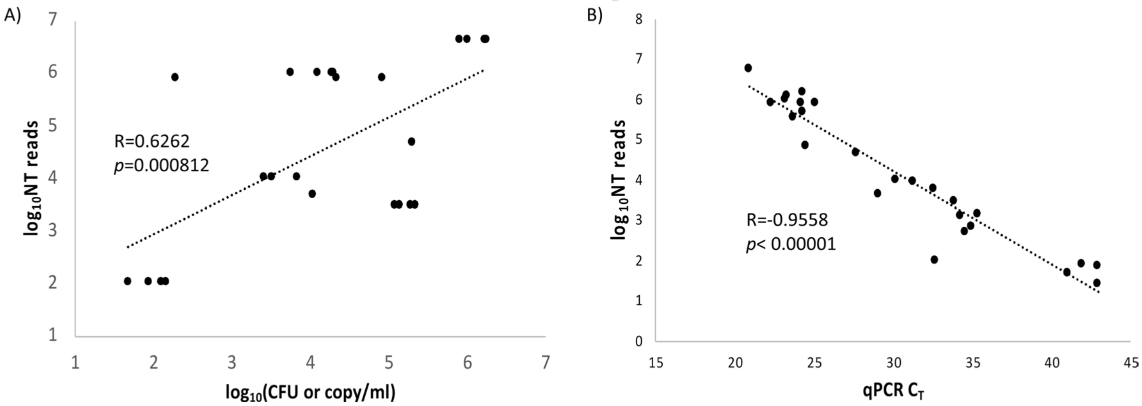

Figure 1. Analytical sensitivity of mNGS to detect pathogens in CSF specimens is comparable to qPCR. CSF specimens (n = 9) negative by standard microbiological methods were spiked with a range of viral and bacterial pathogens at varying concentrations as described in the Materials and Methods or left unspiked and simultaneously tested along with a nuclease free water (NFW) sample by pathogen specific qPCR and by mNGS as described in the Materials and Methods. The approximate titer of pathogens (**A**) or qPCR C$_T$ (**B**) values were plotted against mNGS read counts of pathogens obtained after analysis in IDseq.

Organism	Approximate titer Log(CFU or $TCID_{50}$)/ml	Log_{10}(IDseq NT reads)				Average	SD	CV%
		Experiment 1	Experiment 2	Experiment 3	Experiment 4			
Escherichia coli	4.0 CFU/ml	3.42	3.83	3.52	3.83	3.65	0.21	5.87
Haemophilus influenzae	2.0 CFU/ml	1.68	2.17	1.94	2.11	1.98	0.22	11.06
Neisseria meningitidis	6.6 CFU/ml	5.90	6.24	6.02	6.24	6.10	0.17	2.74
Streptococcus agalactiae	6.0 CFU/ml	4.30	4.11	4.28	3.76	4.11	0.25	6.09
Herpes simplex virus 2	3.5 TCID50/ml	5.08	5.28	5.15	5.34	5.21	0.12	2.21

Table 2. Reproducibility of mNGS assay. *SD* standard deviation, *CV%* coefficient of variance %, *NT* nucleotide, *TCID* tissue culture infectious dose, *CFU* colony forming unit.

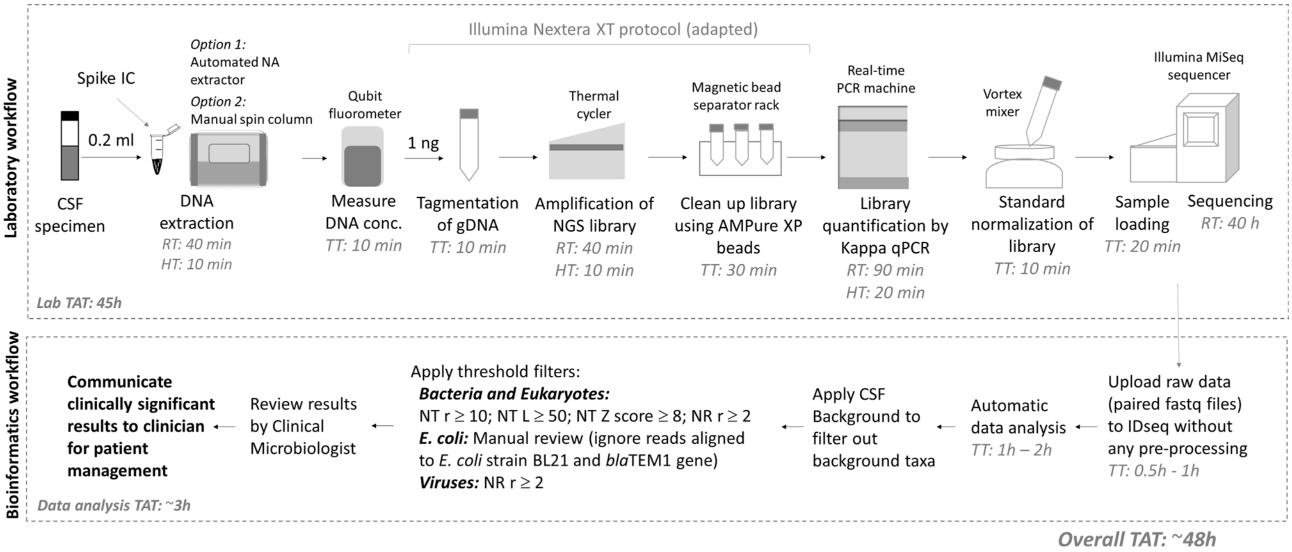
Laboratory workflow
Spike IC
0.2 ml
CSF specimen
Option 1: Automated NA extractor
Option 2: Manual spin column
DNA extraction
RT: 40 min
HT: 10 min
Illumina Nextera XT protocol (adapted)
Qubit fluorometer
Measure DNA conc.
TT: 10 min
1 ng
Tagmentation of gDNA
TT: 10 min
Thermal cycler
Amplification of NGS library
RT: 40 min
HT: 10 min
Magnetic bead separator rack
Clean up library using AMPure XP beads
TT: 30 min
Real-time PCR machine
Library quantification by Kappa qPCR
RT: 90 min
HT: 20 min
Vortex mixer
Standard normalization of library
TT: 10 min
Illumina MiSeq sequencer
Sample loading
TT: 20 min
Sequencing
RT: 40 h
Lab TAT: 45h
Bioinformatics workflow
Upload raw data (paired fastq files) to IDseq without any pre-processing
TT: 0.5h - 1h
Automatic data analysis
TT: 1h – 2h
Apply CSF Background to filter out background taxa
Apply threshold filters:
Bacteria and Eukaryotes:
NT r ≥ 10; NT L ≥ 50; NT Z score ≥ 8; NR r ≥ 2
E. coli: Manual review (ignore reads aligned to E. coli strain BL21 and blaTEM1 gene)
Viruses: NR r ≥ 2
Review results by Clinical Microbiologist
Communicate clinically significant results to clinician for patient management
Data analysis TAT: ~3h
Overall TAT: ~48h

Figure 2. Schematic of mNGS laboratory and bioinformatics workflow. Total nucleic acids (NA) from CSF specimens were extracted after spiking pUC19 plasmid as an internal control (IC) on an automated extraction platform, Qiasymphony (Qiagen). However, any other extraction platform or manual, spin column based methods can also be used for DNA extraction. DNA concertation was determined using a Qubit fluorometer (Thermofisher). NGS library preparation and clean up were performed according to manufacturer's instructions with the exception that 96-well plates were replaced by single tubes or PCR tube strips because single sample was processed. NGS library quantification, normalization, sequencing and data analysis were performed as described in the materials and methods. *TT* total time, *HT* hands-on-time, *RT* reaction time, *TAT* turnaround time. *NT r* nucleotide reads, *NT L* nucleotide alignment length in bp, *NT Z score* nucleotide Z score, *NR r* non-redundant reads.

specific but had poor sensitivity. For example, Metaphlan2 failed to detect low concentrations of *S. pneumoniae* DNA and *H. influenzae* DNA in all samples. Even adenoviral DNA was undetectable despite being strongly positive by qPCR (Supplementary Fig. 3). In order to improve the specificity of IDseq results instead, we first applied a background subtraction method. A background dataset was created from 40 negative CSF samples from the clinical validation study. The background dataset was then applied to the training set (n = 10) and various filters were applied to adjust the sensitivity and specificity of IDseq results based on previous knowledge of spiked organisms and qPCR results. The optimal filter set that cleared all false positive results from the water sample and negative CSF samples were: (i) NT reads ≥ 10; (ii) NT Z score ≥ 8; (iii) aligned base pair length ≥ 50 bp; and (iv) Non redundant (NR) reads ≥ 2. While this approach significantly improved the specificity of mNGS (Supplementary Fig. 4), true positive results for *E. coli* were filtered out because *E. coli* was also a common contaminant in mNGS in our setting. Therefore, we established a manual review process for *E. coli* results. We noted that contaminating *E. coli* DNA in our specimens mapped to *E. coli* strain BL21 or *E. coli* blaTEM1 gene and therefore these results can be ignored. *E. coli* results were only considered as positive if the NGS reads maps to a different strain of *E. coli* as the top hit. A phylogenetic tree of *E. coli* reads from the training set shows that *E. coli* DNA in most negative samples are closely related to either *E. coli* strain BL21 or *E. coli* blaTEM1 gene and all spiked specimens containing true *E. coli* DNA are closely related to NCBI reference genome *E. coli* CFT073. A review of individual sequence alignments in IDseq revealed that the only negative sample (Neg1) that clustered with *E. coli* CFT073 genome also has *E. coli* strain BL21 as its top hit (Supplementary Fig. 5) and the number of sequence reads in this specimen aligned to *E. coli* CFT073 genome was much lower than BL21 (6 aligned reads versus 58 reads). We noted that unlike bacteria and eukaryotes, low count true viral reads were filtered out when the same filter set was applied. Therefore, to identify viral DNA in the test set data, we decided to apply a simple filter of only "NR reads ≥ 2" after background subtraction. All options for applying the background and customized filters are readily available in IDseq and are easy to select with few mouse clicks. The application of these customized filters in conjunction with the manual review for *E. coli* significantly improved the specificity of IDseq returned results without affecting sensitivity. However, the results should be interpreted by a clinical microbiologist to rule out: (i) taxa that merely appear in the results because of sequence homology with the top hit; (ii) clinically insignificant taxa or potential artifacts (such as *Waddlia chondrophila* in Supplemental Fig. 4); and (iii) potential contaminant taxa introduced during specimen collection or handling and processing of specimens. In cases where more than one taxa is returned after applying custom filters on IDseq, the clinical microbiologist should carefully review the results. Organisms with the highest number of reads mapped to the reference genomes should be considered to be positive unless deemed clinically irrelevant by the clinical microbiologist. Taxa that are known to be closely related to the top scoring organism should be ignored. A simplified bioinformatics workflow for clinical validation of mNGS is shown in Fig. 2.

Clinical validation of mNGS for pathogen detection in CSF.

For clinical validation, previously saved residual CSF specimens (Test set; n = 74) that were submitted for microbiological assessment were selected for mNGS analysis (Table 3). The samples were collected from patients with suspected CNS infections. The average age of patients was 3.3 years. 35% patients were neonates (≤ 28 days) and 68% were male. The neonatal and pediatric intensive care units and the emergency department accounted for about 26% and 15% specimens, respectively. 55% specimens came from other in-patient units including neuroscience and neurosurgery, respiratory and cardiac services and renal, endocrine and metabolic units. A review of patient's laboratory data indicated that about 46% CSF specimens had abnormal white blood cell (WBC) counts and red blood cells (RBC) were detected in most of the specimens. The proportion of specimens with very high WBC (≥ 1,000 per mm^3) and RBC (≥ 400 per mm^3) counts were 14% and 27%, respectively. Specimens considered to be "bloody taps" (RBC/WBC > 500) were excluded. By Gram staining, about 23% and 18% specimens had very high (4 +) WBC and RBC scores, respectively. By culture, 11 samples (15%) were positive and when combined with the PCR results 28 (39%) samples were positive for a bacterial, viral or fungal pathogen. A total of 9 samples were positive for enterovirus, which is an RNA virus. Because RNA viruses are not expected to be detected by the DNA-specific protocol described in this study, these samples were considered as negative samples for validation purposes (Supplementary Table 2).

The DNA concentration in CSF specimens ranged from 0.14 to 19.9 ng/µl with an average of 1.3 ng/µl (Supplementary Table 1). Because the Illumina Nextera XT protocol requires 1 ng DNA in 5 µl volume, 11 specimens had lower DNA concentration than the minimally required concentration. For these specimens, 5 µl of the undiluted DNA extracts were used for library preparation irrespective of their concentration. Other specimen extracts were diluted to 1 ng DNA as the starting material for library preparation. The NGS library concentration based on Kappa qPCR ranged from 0.4 to 150 nM with an average of 19.9 nM. The Illumina MiSeq protocol requires at least 2 nM DNA but 12 specimens had lower than the minimum library concentration. For these specimens, 5 µl of the undiluted libraries were processed for sequencing, while other libraries were diluted as required. A total of about 5 h, with approximately 2 h of hands-on time, was required for DNA extraction, NGS library preparation, quantification, normalization and sample loading. Sequencing on the MiSeq required about 40 h per run (Fig. 2). The total sequence read output ranged from 3,757,836 to 44,424,176 with an average of 23,463,311. The average sequence read outputs obtained from specimens with < 0.2 ng/µl DNA or with libraries with < 2 nM concentration (25,201,237 and 20,188,913, respectively) were not significantly different from the overall average sequence read outputs. Non-host reads ranged from 574 to 2,430,956 reads with an average of 56,179 reads, accounting for an average of 0.3% of total reads. The average run time for data analysis was 1.22 h. Data analysis for 90% and 74% of specimens were completed in < 2 and < 1 h, respectively.

For pathogen detection, mNGS data were analyzed using the IDseq pipeline according to the standard operating procedures described in Fig. 2. After applying customized filter-sets, the results were reviewed and called by a clinical microbiologist in a blinded manner, and compared with culture and PCR results (Supplementary Table 2). Results were classified as true positive, true negative, false positive and false negative based on conventional tests results that were originally reported for patient management. Discrepant results and some results with very low mNGS reads were also confirmed by pathogen specific PCR or Sanger sequencing during the course of this study. Samples CW005 and CW322 were considered negative for the presence of a CNS pathogen by conventional tests, because they were originally reported as potential contaminants. *Staphylococcus epidermidis* in sample CW005 grew in enrichment broth only. On the other hand, one colony each of *Staphylococcus epidermidis* and *Roseomonas* spp. grew from sample CW322, but this sample was also strongly positive for *Neisseria menigitidis* by qPCR. A total of 21 samples were called positive for the presence of a pathogen by mNGS (Fig. 3). mNGS identified three additional pathogens that were not detected by conventional methods. *S. agalactiae* was detected in two specimens from patients aged 0.09 and 0.03 years, respectively. The organisms were not recovered by culture, but gram-positive cocci were noted on Gram stain and were positive by confirmatory qPCR subsequently during the course of this study. In specimen CW060, mNGS detected few sequence reads that aligned to an *Acinetobacter* sp. TGL-Y2 plasmid, which was deemed clinically insignificant. A few reads of a *Streptococcus parasanguinis* were also detected in this specimen. However, a PCR targeting the bacterial 16S rRNA gene returned negative results for both specimens CW005 and CW060 (data not shown). *Candida tropicalis* DNA in sample CW101 was also confirmed by ITS sequencing because of low read counts. Finally, low level HSV2 DNA was detected

Characteristics	No. of samples	% of total
Age		
All	74	100.0
0–28 days	26	35.1
28 days–3 months	12	16.2
3 months–5 years	19	25.7
5–18 years	16	21.6
> 18 years	1	1.4
Gender		
Male	50	67.6
Female	24	32.4
Location		
All	74	100.0
Intensive care	19	25.7
Emergency	11	14.9
Inpatient medical and surgical units	41	55.4
Others	3	4.1
CSF cell count and differentiation		
All	74	100.0
WBC (≥ 5 per mm^3)	34	45.9
WBC (≥ 1,000 per mm^3)	10	13.5
RBC (≥ 1)	67	90.5
RBC (≥ 400)	20	27.0
Neutrophil (≥ 50%)	17	23.0
Lymphocyte (≥ 50%)	19	25.7
Monocyte (≥ 50%)	20	27.0
CSF chemistry		0.0
Protein (≥ 1.5 g per L)	17	23.0
Glucose (< 1.7 mmol/L)	9	12.2
Microbiology		
All	74	100.0
GS, WBC (1 +)	54	73.0
GS, WBC (4 +)	17	23.0
GS, RBC (1 +)	31	41.9
GS, RBC (4 +)	13	17.6
Culture positive	11	14.9
Culture and/or PCR positive	28	37.8
Culture and/or PCR positive except enterovirus	19	25.7

Table 3. Specimen and patient characteristics. *WBC* white blood cells, *RBC* red blood cells, *GS* Gram staining.

in sample CW322, which was strongly positive for *N. meningitidis*. Unfortunately, the presence of HSV2 DNA could not be further verified by PCR or any other methods because no more specimen was available for analysis.

Interestingly, our *E. coli* approach established based on training dataset correctly differentiated true positive results from false positive, background *E. coli* DNA, when applied to the validation dataset. A phylogenetic tree built based on the validation dataset clearly separated out *E. coli* DNA in 2 specimens that were related to the *E. coli* CFT073 genome from those that are related to the BL21 strain, and were interpreted as positive results (Supplementary Fig. 6). It was noted that *E. coli* DNA in 69 out of 74 specimens were related to BL21 strains. The remaining 3 specimens were related to *E. coli* reference strain C9. However, all of these specimens were strongly positive for *N. meningitidis* and were therefore ignored. Among the viral taxa identified, Chimpanzee anellovirus and Torque Teno mini virus were deemed insignificant by the clinical microbiologist.

Overall, the sensitivity, specificity and accuracy of mNGS results on 74 CSF specimens with the customized bioinformatic approach in IDseq were 100%, 95% and 96%, respectively (Table 4). For comparison, all the mNGS data were also analyzed by the Metaphlan2 pipeline and the results were reviewed by a clinical microbiologist. All *E. coli* positive results, common environmental contaminants such as *Ralstonia picketii* and clinically non-relevant taxa were interpreted as negative results. The sensitivity, specificity and accuracy of mNGS results using Metaphlan2 were 58%, 96% and 86% respectively.

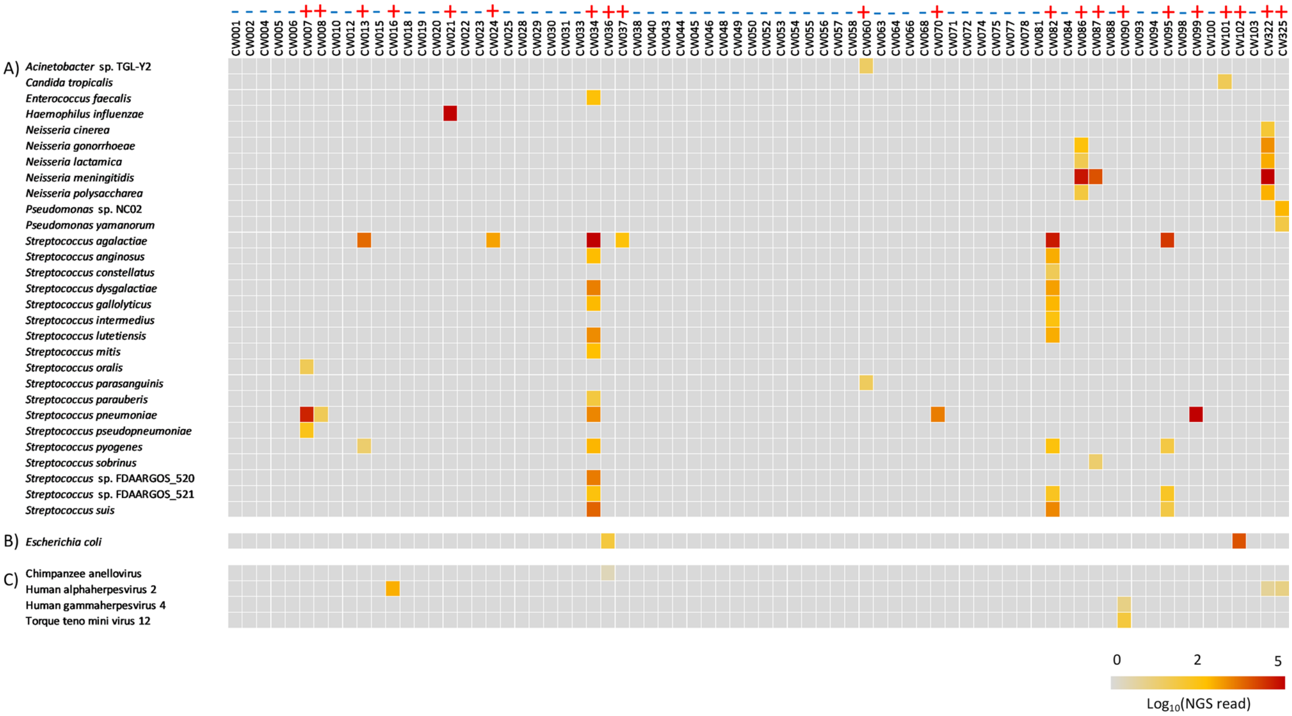

Figure 3. Heatmap of taxa identified in the validation set of specimens with IDseq after applying CSF background and customized threshold filters. mNGS was performed and data were analyzed as described in Fig. 2. Heatmaps were generated based on $\log_{10}$[NT total reads]. Different filter sets were applied for bacteria and eukaryotes versus viruses and *E. coli* reads were manually derived from IDseq after manual review of data. All positive and negative results by mNGS are shown by '+' and '–' signs, respectively.

Statistics	Metaphlan2	IDseq
Total number of results	74	74
True positive	11	18
True negative	53	53
False positive	2	3
False negative	8	0
Sensitivity	58% (95% CI 34–80%)	100% (95% CI 81–100%)
Specificity	96% (95% CI 87–100%)	95% (94% CI 85–99%)
Accuracy	86% (95% CI 77–93%)	96% (95% CI 87–99%)

Table 4. Diagnostic performance of mNGS assay compared to conventional methods for CNS infections.

Discussion

The potential for the application of NGS based, clinical metagenomics approach for the diagnosis of infectious diseases has been recognized for a few years now[8, 9, 15]. Currently, initiatives are being taken to standardize the approach for routine application in the clinical microbiology laboratories[20, 22, 23]. Because of the high cost and substantial expertise and labor involved, mNGS approach appears suitable only for specific applications in clinical microbiology such as for detection of pathogens in CSF for the diagnosis of meningitis and encephalitis because these infections are potentially life threatening and also because of the fact that a wide range of pathogens can cause such infections that are not routinely detected by standard methods. The present study was performed to establish and validate an mNGS approach for pathogen detection in CSF for the diagnosis of CNS infections that can be implemented in acute care hospital laboratories. The protocol has been developed for a low throughput setting so that single specimens can be processed without waiting for additional specimens for batching. Laboratory workflow and the data analysis workflow have been designed to be relatively simpler and faster compared to previously reported approaches. The procedure does not require huge capital investment or large-scale instrumentation and automation, and can be performed by molecular microbiology technologists with a minimum of extra training. As with current culture methods however, the results must be reviewed by a clinical microbiologist for clinical correlation and to rule out false positive results arising from potential contaminants or clinically insignificant taxa.

Among the various platforms of Illumina that offers NGS at various depths and scales, the benchtop sequencer MiSeq was chosen because the equipment cost is relatively lower and the platform is more suitable for low throughput applications. The MiSeq platform generates enough data and sequence reads so that low level pathogen DNA can be detected. Our analytical sensitivity data shows that the MiSeq-based approach was able to detect low level pathogen DNA with qPCR C_T 35 or higher. Even spiked *H. influenzae* that was undetectable by qPCR was detected by mNGS. Given the fact that PCR is widely considered to be one of the most sensitive method for pathogen detection, an equivalent or superior sensitivity to qPCR indicates that mNGS approach can be applied for the diagnosis of infectious diseases. In the clinical validation, we noted that despite minimal quality control and high variability in extracted DNA concentration, NGS library concentration, sequence read output, host versus non-host reads and IC reads, the mNGS approach was capable of identifying all pathogens identified by conventional methods except a few pathogens that were originally reported as contaminants. On the other hand, the fact that mNGS was able to identify *S. agalactiae* in two specimens that did not grow in culture suggests that mNGS approach is a powerful method for identification of fastidious pathogens or those that have failed to grow in culture because of prior antibiotic treatment.

Because NGS generates massive amounts of sequence data, interpretation of this data for clinical use is challenging. In order to develop a practical bioinformatics approach, with high sensitivity and specificity for pathogen detection in CSF, we tested various existing bioinformatics tools including Metaphlan2[27] and IDseq[28] using our training set. Metaphlan2 performs taxonomic assignment of metagenomic shotgun sequencing data at the species level using a unique database of clade specific marker genes as the reference database[27]. Metaphlan2 is performed in command line, Linux or MacOS environments and therefore requires some bioinformatics expertise and experience working in a command line environment. IDseq is an online platform for metagenomic sequence analysis, where the user uploads raw sequence data to the platform and the rest of the procedure, including adapter trimming, data quality control, host DNA subtraction and alignment to both NCBI nucleotide (NT) database and non-redundant (NR) database is performed automatically.

When Metaphan2 was applied to our training and validation dataset, the specificity was high but the sensitivity was poor. For example, in the validation dataset, Metaphlan2 missed 8 of the 19 positive results by standard methods. On the other hand, when IDSeq pipeline was applied to our training data set, all spiked pathogens were detected with high sensitivity but the metagenomic sequence analysis returned a long list of bacterial, viral and eukaryotic taxa making it very difficult to differentiate true positive results from background taxa and/or potential contaminants, particularly when the target pathogen was present at low quantities (Supplementary Fig. 1). The IDseq platform, however, offered us with a wide range of options to utilize additional statistical analyses and data filters to improve the sensitivity and specificity of the methods and allow for easier interpretation of data. First, we created a background dataset using data from 40 known negative samples. After background subtraction, while IDseq ranks taxa based on an aggregate score calculated from NT/NR 'z scores' and 'reads per million' (rpm), this method alone was not sufficient to delineate specific results from the list of potential pathogens and non-specific taxa. We therefore applied additional filters and manual review process as described in the results section in order to improve the interpretability of results. A manual review process was necessary for *E. coli*, because the bacteria is a significant CNS pathogen but is also a common contaminant in our setting. We noted that a particular strain of this bacteria, *E. coli* BL21, is present in almost all samples. *E. coli* BL21 is a laboratory strain widely known for use in plasmid preparation, cloning and recombinant enzyme production. We speculate that DNA from this bacterial strain may come from reagent contamination or may have been co-extracted with the specimens because of the plasmid internal control that we spiked into the specimens before extraction. Interestingly, when this approach was applied to the validation data set, the returned results were highly accurate, sensitive and specific (≥ 95%) compared to the standard methods. Clinical microbiologist input is critical for accurate interpretation of mNGS data to prevent reporting of potentially contaminating taxa or clinically irrelevant taxa. Review of microbiological results by the clinical microbiologists before they are reported for patient management is a routine practice in diagnostic microbiology laboratories and therefore not exceptional for mNGS.

A limitation of our current approach is that our method is not completely unbiased for detection of CNS pathogens because we did not incorporate methods for detecting RNA viruses. A combined method for detection of both DNA and RNA targets by mNGS would expand the target pathogen range of the assay but would increase the cost of testing and increase the complexity of the procedure. Another limitation is that most pathogens detected by mNGS in our study were relatively common bacterial and viral pathogens, and no uncommon CNS pathogens were identified in specimens that were negative by standard methods. These results perhaps reflects the nature of our pediatric population in British Columbia, Canada. However, the high sensitivity and specificity of our approach suggests the potential for detecting more unusual pathogens in a more diverse population. Similar studies conducted in populations or geographic regions with higher rates of unusual CNS pathogens may demonstrate further benefits of applying mNGS for the diagnosis of CNS infections. We are currently undertaking a study to investigate the clinical impact of implementing mNGS for prospective pathogen detection in selected CSF specimens in a pediatric population in Qatar.

In conclusion, we have developed a clinical metagenomic diagnostic approach for CNS infections which has demonstrated superior accuracy, sensitivity and specificity, compared to previously reported methods. The method offers relatively faster turnaround time, minimal hands-on time and can be easily implemented in an acute care diagnostic microbiology laboratory, with a moderate level of molecular expertise, and with modest capital investment. Currently, very few reference laboratories of the world offer clinical metagenomics based diagnostic services for infectious diseases. These laboratories are highly specialized and are equipped with large, production scale sequencing platforms and massive computational servers. The cost of sequencing is reduced through sample batching, but this, along with time required for sample shipping, increases the turnaround time, which can be detrimental for managing CNS infections. Therefore, the method described here, designed for use in a hospital acute care setting, should be of immediate benefit to patients with undiagnosed CNS infections.

Materials and methods

Bacterial and viral strains. Bacterial strains used for spiking in this study were *Escherichia coli* (American Type Culture Collection [ATCC] 25922), *Streptococcus pneumoniae* (ATCC49619), *Streptococcus agalactiae* (ATCC12386), *Haemophilus influenzae* (ATCC10211) and *Neisseria meningitidis* (ATCC 13090) and viral strains include Herpes Simplex Virus 2 (HSV2) (ATCC VR-540) and Human Adenovirus type 7 (ATCC VR-7). *E. coli, S. pneumoniae* and *S. agalactiae* were grown on blood agar plates (Oxoid) overnight at 37 °C in a 5% CO_2 atmosphere. *H. influenzae* and *N. meningitidis* were grown on chocolate agar plates (Oxoid) under the same conditions. Viral stocks were maintained in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% DMSO at − 80 °C.

Specimens. For test optimization and spiking, 9 CSF specimens submitted to the Microbiology and Virology laboratory of BC Children's Hospital for culture and PCR testing between August 2013 and July 2014 were used. All CSF specimens were negative by culture and PCR for all pathogens that were used for spiking in this study. Following standard testing, residual specimens were kept at 4 °C until processed. For clinical validation, 74 CSF specimens that were collected from August 2015 to December 2017 from children admitted to BC Children's Hospital with suspected CNS infections were used in a retrospective manner. Residual specimens were saved at -80 °C after standard testing. To maintain patient anonymity all patient identifiers were removed by hospital staff unaware of the current study results. Patient data including age and gender, ordering location, collection date, and laboratory data including CSF cell count and chemistry and Gram staining, culture and PCR results were recorded against a study specific identifier in a spreadsheet. Ethics approval for the study was obtained from the Research Ethics Boards (REB) of University of British Columbia, BC, Canada and Sidra Medicine, Doha, Qatar. A waiver of informed consent was requested and was approved by REBs of both institutions. All methods were performed in accordance with the relevant guidelines and regulations.

Establishment of a training set. Five out of nine negative CSF specimens were spiked with a range of bacterial and viral species. The remaining 4 were left un-spiked to serve as negative control. In addition, an un-spiked nuclease free water (NFW) sample was also included in the training set and processed simultaneously. For spiking, bacterial suspensions were freshly prepared in phosphate buffered saline (PBS) to a turbidity equivalent to a 0.5 McFarland standard and further diluted 10, 100 and 1,000-fold in PBS, as required. *N. meningitidis* (4.3×10^8 CFU/ml), HSV2 (2.8×10^5 $TCID_{50}$/ml) and adenovirus (2.8×10^6 $TCID_{50}$/ml) were directly added from cultured stocks. Bacterial and viral preparations were spiked into 0.5 ml of CSF specimens to give approximate final titers shown in Table 1, and vortexed for 10 s before extraction of DNA and analysis by qPCR and mNGS as described below.

qPCR and DNA sequencing. Residual clinical samples were extracted on a QIAsymphony instrument (Qiagen) using the DSP Virus/Pathogen Mini kit. Total nucleic acids from spiked or unspiked specimens were extracted and analyzed by qPCR assays for various pathogens as described previously[29] and DNA concentration was measured in a Qubit 2.0 fluorometer using the Qubit dsDNA HS assay kit (Thermo Fisher Scientific, Inc.). To serve as an internal control for extraction, qPCR and NGS, a purified plasmid DNA pUC19 was spiked to each specimen prior to extraction at a final concentration of 1.4×10^5 copies/ml. NGS libraries were prepared from 1 ng of extracted DNA using Nextera®XT DNA Sample Preparation Kit (Illumina), and sequencing was performed on an Illumina MiSeq sequencer using MiSeq Reagent Kit v2 (500-cycles) (Illumina) as described previously[29]. The concentration of prepared NGS libraries were determined by Kappa qPCR according to manufacturer's instructions (Roche). Sequencing was performed in 3 different facilities: a) McGill University and Génome Québec Innovation Centre, Montréal (Québec), Canada b) Sidra Medicine, Doha, Qatar and c) Alliance Global Middle East, Beirut, Lebanon. Specimens were sequenced irrespective of the quality of NGS libraries. Confirmatory qPCR for *S. agalactiae* were performed as described previously[29]. Bacterial 16S rRNA PCR was performed as described previously[30].

Bioinformatics. Raw sequence data were analyzed either by Metaphlan2 pipeline[27] or uploaded to https://idseq.net for automated metagenomic analysis without any pre-processing of data. IC reads were obtained by mapping raw sequence reads to pUC19 plasmid sequence using Bowtie2 plugin in Geneious 11.1.5 software.

Statistical analysis. The linear correlation of NGS data with that of PCR results was determined by calculating Pearson product-moment correlation coefficient in excel followed by determining the significance level at $p < 0.05$. Diagnostic sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and accuracy (concordance) were calculated using an online, diagnostic test evaluation calculator, and associated 95% confidence intervals (CI) were calculated by the Clopper-Pearson interval or exact method using the same calculator[31].

Data availability

The non-host reads from simulated and patient CSF samples tested in this study are available in https://idseq.net/ under the project name 'CSF_metagenomics'.

Received: 6 November 2019; Accepted: 16 June 2020
Published online: 08 July 2020

References

1. Hasbun, R. The acute aseptic meningitis syndrome. *Curr Infect Dis Rep* **2**, 345–351 (2000).
2. Logan, S. A. & MacMahon, E. Viral meningitis. *BMJ* **336**, 36–40. https://doi.org/10.1136/bmj.39409.673657.AE (2008).
3. Tunkel, A. R., Beek, D. & Scheld, W. M. In *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases* Vol. 1 (eds Bennett, J. E. *et al.*) 1097–1137 (Elsevier Inc., Amsterdam, 2015).
4. Bennett, J. E. In *Mandell, Douglas, and Bennet's Principles and Practice of Infectious Diseases* (eds Bennett, J. E. *et al.*) 1138–1143 (Elsevier, Philadelphia, 2015).
5. Beckham, J. T. & Tyler, K. L. In *Mandell, Douglas, and Bennet's Principles and Practice of Infectious Diseases* Vol. 1 (eds Bennett, J. E. *et al.*) 1144–1163 (Elsevier, Philadelphia, 2015).
6. Polage, C. R. & Cohen, S. H. State-of-the-art microbiologic testing for community-acquired meningitis and encephalitis. *J. Clin. Microbiol.* **54**, 1197–1202. https://doi.org/10.1128/JCM.00289-16 (2016).
7. Takhar, S. S., Ting, S. A., Camargo, C. A. & Pallin, D. J. U. S. emergency department visits for meningitis, 1993–2008. *Acad. Emerg. Med.* **19**, 632–639. https://doi.org/10.1111/j.1553-2712.2012.01377.x (2012).
8. Goldberg, B., Sichtig, H., Geyer, C., Ledeboer, N. & Weinstock, G. M. Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. *MBio* **6**, e01888-e11815. https://doi.org/10.1128/mBio.01888-15 (2015).
9. Kwong, J. C., McCallum, N., Sintchenko, V. & Howden, B. P. Whole genome sequencing in clinical and public health microbiology. *Pathology* **47**, 199–210. https://doi.org/10.1097/PAT.0000000000000235 (2015).
10. Tang, P. & Chiu, C. Metagenomics for the discovery of novel human viruses. *Future Microbiol.* **5**, 177–189. https://doi.org/10.2217/fmb.09.120 (2010).
11. Bogaert, D. *et al.* Variability and diversity of nasopharyngeal microbiota in children: a metagenomic analysis. *PLoS ONE* **6**, e17035. https://doi.org/10.1371/journal.pone.0017035 (2011).
12. Barzon, L., Lavezzo, E., Militello, V., Toppo, S. & Palu, G. Applications of next-generation sequencing technologies to diagnostic virology. *Int J Mol Sci* **12**, 7861–7884. https://doi.org/10.3390/ijms12117861 (2011).
13. Yang, J. *et al.* Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *J. Clin. Microbiol.* **49**, 3463–3469. https://doi.org/10.1128/JCM.00273-11 (2011).
14. Nakamura, S. *et al.* Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS ONE* **4**, e4219. https://doi.org/10.1371/journal.pone.0004219 (2009).
15. Wilson, M. R. *et al.* Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N. Engl. J. Med.* **370**, 2408–2417. https://doi.org/10.1056/NEJMoa1401268 (2014).
16. Consortium, H. M. P. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214. https://doi.org/10.1038/nature11234 (2012).
17. Liu, X. *et al.* A tentative tamdy orthonairovirus related to febrile illness in Northwestern China. *Clin. Infect. Dis.* https://doi.org/10.1093/cid/ciz602 (2019).
18. Chiu, C. Y. & Miller, S. A. Clinical metagenomics. *Nat. Rev. Genet.* **20**, 341–355. https://doi.org/10.1038/s41576-019-0113-7 (2019).
19. Gu, W., Miller, S. & Chiu, C. Y. Clinical metagenomic next-generation sequencing for pathogen detection. *Annu. Rev. Pathol.* **14**, 319–338. https://doi.org/10.1146/annurev-pathmechdis-012418-012751 (2019).
20. Miller, S. *et al.* Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome Res.* **29**, 831–842. https://doi.org/10.1101/gr.238170.118 (2019).
21. Schlaberg, R. *et al.* Validation of metagenomic next-generation sequencing tests for universal pathogen detection. *Arch. Pathol. Lab. Med.* **141**, 776–786. https://doi.org/10.5858/arpa.2016-0539-RA (2017).
22. Wilson, M. R. *et al.* Clinical metagenomic sequencing for diagnosis of meningitis and encephalitis. *N. Engl. J. Med.* **380**, 2327–2340. https://doi.org/10.1056/NEJMoa1803396 (2019).
23. Hasman, H. *et al.* Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *J. Clin. Microbiol.* **52**, 139–146. https://doi.org/10.1128/JCM.02452-13 (2014).
24. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624. https://doi.org/10.1038/ismej.2012.8 (2012).
25. Ramesh, A. *et al.* Metagenomic next-generation sequencing of samples from pediatric febrile illness in Tororo, Uganda. *PLoS ONE* **14**, e0218318. https://doi.org/10.1371/journal.pone.0218318 (2019).
26. Saha, S. *et al.* Unbiased metagenomic sequencing for pediatric meningitis in Bangladesh reveals neuroinvasive chikungunya virus outbreak and other unrealized pathogens. *mBio* **10**(6), e02877-19. https://doi.org/10.1128/mBio.02877-19 (2019).
27. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814. https://doi.org/10.1038/nmeth.2066 (2012).
28. Bouncy, C. D. *IDseq: An Open Source Platform for Infectious Disease Detectives*, https://medium.com/czi-technology/a-platform-for-infectious-disease-detectives-253753026fe8 (2018).
29. Hasan, M. R. *et al.* Depletion of human DNA in spiked clinical specimens for improvement of sensitivity of pathogen detection by next-generation sequencing. *J. Clin. Microbiol.* **54**, 919–927. https://doi.org/10.1128/JCM.03050-15 (2016).
30. El-Nemr, I. M. *et al.* Application of MALDI biotyper system for rapid identification of bacteria isolated from a fresh produce market. *Curr. Microbiol.* **76**, 290–296. https://doi.org/10.1007/s00284-018-01624-1 (2019).
31. MedCalc. *Diagnostic test evaluation calculator,* https://www.medcalc.org/calc/diagnostic_test.php, Accessed on October 01, 2018, https://www.medcalc.org/calc/diagnostic_test.php (2018).

Acknowledgements

The study was supported by British Columbia Children's Hospital Foundation, Canada (Grant no. KRZ28061 to M.R.H.) and Sidra Medicine, Qatar (Grant no. SIRF_200026 to M.R.H.). We are grateful to the IDseq team for providing technical support for data upload and analysis.

Author contributions

P.T., R.T. and M.R.H. conceived the idea. M.R.H. designed the study, performed data analysis and wrote the manuscript. S.S. and K.M.T. performed the laboratory work. P.T. obtained specimens and associated clinical and laboratory data from the clinical cohort. P.T. provided critical inputs on experimental approach and data analysis. A.P.L. and M.J. provided input on data interpretation. All authors contributed to the revision of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-68159-z.

Correspondence and requests for materials should be addressed to M.R.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.