Targeted metagenomic sequencing enhances the identification of pathogens associated with acute infection.

Cyndi Goh[1]*, Tanya Golubchik[1,2]*, M Azim Ansari[1,3]*, Mariateresa de Cesare[1,2], Amy Trebes[1], Ivo Elliott[1,4], David Bonsall[1,2], Paolo Piazza[1], Anthony Brown[3], Hubert Slawinski[1], Natalie Martin[5], Sylviane Defres[6], Mike J Griffiths[6,7], James E Bray[8], Martin C Maiden[8], Paula Hutton[9], Charles J Hinds[10], Tom Solomon[6,11], Ellie Barnes[3], Andrew J Pollard[5,12], Manish Sadarangani[5+], Julian C Knight[1+], Rory Bowden[1+**]

[1] Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK

[2] Big Data Institute, University of Oxford, Oxford, UK

[3] Peter Medawar Building for Pathogen Research, University of Oxford, Oxford, UK

[4] Lao-Oxford-Mahosot Hospital–Wellcome Trust Research Unit, Microbiology Laboratory, Mahosot Hospital, Vientiane, Lao People's Democratic Republic

[5] Oxford Vaccine Group, Department of Paediatrics, University of Oxford, Oxford UK

[6] Institute of Infection & Global Health, University of Liverpool, Liverpool, UK

[7] National Institute for Health Research Health Protection Research Unit in Emerging and Zoonotic Infections, University of Liverpool, Liverpool, UK

[8] Department of Zoology, University of Oxford, Oxford, UK

[9] Adult Intensive Care Unit, John Radcliffe Hospital, Oxford, UK

[10] William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University, London, UK

[11] Walton Centre NHS Foundation Trust, Liverpool, UK

[12] NIHR Oxford Biomedical Research Centre, Oxford UK

* These authors contributed equally to this work.

[+] Jointly directed the study

** Corresponding author: rbowden@well.ox.ac.uk Wellcome Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK

Abstract

The routine identification of pathogens during infection remains challenging because it relies on multiple modalities such as culture and nucleic acid amplification and tests that tend to be specific for very few of an enormous number of possible infectious agents. Metagenomics promises single-test identification, but shotgun sequencing remains unwieldy and expensive or in many cases insufficiently sensitive to detect the amount of pathogen material in a clinical sample. Here we present the validation and application of *Castanet*, a method for metagenomic sequencing with enrichment that exploits clinical knowledge to construct a broad panel of relevant organisms for detection at low cost with sensitivity comparable to PCR. *Castanet* targets both DNA and RNA, works with small sample volumes, and can be implemented in a high-throughput diagnostic setting. We used *Castanet* to analyse plasma samples from 573 patients from the GAinS sepsis cohort and CSF samples from 243 patients from the ChiMES meningitis cohort that had been evaluated using standard clinical microbiology methods, identifying relevant pathogens in many cases where no pathogen had previously been detected. *Castanet* is intended for use in defining the distribution of pathogens in samples, diseases and populations, for large-scale clinical studies and for verifying the performance of routine testing regimens. By providing sequence as output, *Castanet* combines pathogen identification directly with subtyping and phylo-epidemiology.

Introduction

Many cases of acute infection, including meningitis and sepsis, are managed in the absence of a specific pathogen diagnosis in both high- and low-resource settings[1-5]. The primacy of empirical therapy reflects clinical urgency combined with the complexity of the testing landscape for acute conditions, a lack of assay sensitivity, and the time required to run a series of tests. A large number of potentially causative organisms must be considered, each with its own biology and consequently set of diagnostic tests. Even if all possible tests could be performed quickly enough to inform patient care, test sensitivity is limited by the low pathogen levels in many specimen types and by the progressive depletion of sample volume by each subsequent test. No single test is perfect: traditional culture methods are broadly applicable for bacteria but are slow and biased in favour of culturable and fast-growing organisms. Molecular tests such as organism-specific PCR are fast and specific, but risk false-negative results where variation among pathogen genomes leads to failure to detect the organism[6].

Strategies to reduce the number of sequential tests and speed up pathogen detection have included the development of multi-pathogen assays such as multiplex PCRs[7-10], panels of serological tests[11 12], 16S DNA sequencing[13 14], and most recently, metagenomic sequencing[15 16]. Diagnostic metagenomics – that is, deep sequencing of total nucleic acid directly from a clinical sample – is the only currently available technique that offers the promise of detecting the complete pathogen composition of a sample in a single assay, as well as information on virulence and antimicrobial susceptibility[15]. In practice however, diagnostic metagenomic applications are limited to sample types with high pathogen abundance[17 18]. In most readily available clinical samples, such as blood, the vast majority of sequences originate from the human host, since a single human cell contains as much nucleic acid as a thousand bacteria or up to a million viruses. In addition, material from commensal and contaminant organisms and kit reagents is essentially ubiquitous[19], so a means is needed to distinguish pathogen sequences from background.

To be useful for routine clinical applications, metagenomics requires laboratory techniques that increase the relative yield of sequences that come from the pathogen(s) of interest. In efforts to harness the power of metagenomics for low-abundance samples, various purification or enrichment procedures have been employed to increase yield. For bacterial pathogens, the most common first step is culture, which can bias the result in favour of contaminating, culturable or fast-growing organisms. For viral pathogens and some bacteria, yield-maximising methods include low-speed centrifugation and/or filtration to remove host cells, sample treatment with nucleases to digest nucleic acid not protected within cells or virions[20 21] and high-speed gradient centrifugation to concentrate virus particles (for review, see Duhaime and Sullivan 2012)[22]. Each of these procedures reduces throughput and may bring bias[23].

An alternative and widely applicable approach is targeted enrichment of a subset of sequences in a metagenomic sample using nucleic acid probes, or “baits”. Targeted enrichment works at a range of scales from individual genes to the 50-60+ Mb of sequences defining the human exome[24-27], and can be applied either to unprocessed clinical samples or to sequencing libraries. We previously demonstrated that enrichment is robust to substantial local divergence between probe and target sequences, making it possible to design multi-pathogen probe sets that simultaneously capture a diverse set of clinically relevant pathogen sequences[28]. Targeted enrichment for single or larger numbers of targeted organisms has been demonstrated in the study of viruses[28-32] and bacteria[33 34]. However, significant practical hurdles remain in the application of sequence enrichment as a routine and comprehensive multi-pathogen test, including the development of efficient workflows, the need to balance the opposing considerations of the breadth of species detected and the cost and complexity of large probe panels[32 34 35] and the requirement for bespoke analytical pathways.

Here we describe *Castanet*, a probe-based targeted enrichment methodology that can detect an arbitrarily wide selection of clinically important bacterial and viral pathogens, and interrogates both DNA and RNA sequences in the same workflow. *Castanet* is based on a clinically informed probe panel, designed to replace and augment first-line diagnostic tests for a range of acute infections, and is suitable for use in high-throughput settings. We validate and benchmark method performance both on a standard reference containing a prequantified mixture of known viruses[36], and on clinical samples with well-quantified pathogen load. We then demonstrate the application of our method in clinical settings by using it to detect pathogens in patient samples from sepsis and meningitis, two clinically important, life-threatening syndromes where the available samples are notoriously low in pathogen load, and where conventional diagnostic testing often fails to find an infectious cause[15 37].

Results

Design of probes for an extensive panel of disease-relevant pathogens

We compiled a list of viral and bacterial pathogens relevant to paediatric meningitis and adult sepsis from community-acquired pneumonia (CAP) in the UK (**Table 1**), adding several pathogens of current interest that were less likely to be present in UK samples. Considering the number of distinct entries on our list (116, from 17 virus families and 35 bacterial species), we inferred that any other organisms, including relevant fungal or parasite pathogens, would necessarily comprise rare causes of meningitis or pneumonia and sepsis.

We targeted similar lengths of genomic sequence for each pathogen to achieve comparable assay sensitivity, optimising the breadth of pathogens we could target and avoiding bias in favour of larger genomes. For each of the viruses, we downloaded from NCBI RefSeq the full set of complete genomes available at 1st August 2015. Except for each of the included herpesviruses, whose genomes exceed 100 kbp and where we chose a low-diversity region of ~20kb, for each virus we constructed a whole-genome alignment using MAFFT[38] from which to design probes. For bacterial species, we took advantage of the ribosomal multilocus sequence typing (rMLST) scheme, which targets 53 genes encoding ribosomal proteins present in all bacteria and resolves bacteria to a sub-species level[39], extracting sequences from the rMLST database on 11 December 2015.

In previous work we observed that sequence capture efficiency is preserved even when probe and target sequences diverge substantially, and that exploiting sequence similarity to avoid redundancy can improve the efficiency of probe design without sacrificing performance[28]. Accordingly, for each sequence alignment we constructed a UPGMA tree using pairwise Hamming distances, within which we identified clusters such that all sequences were less than 5% divergent from one another. This process defined a set of 5.86×10^6 bases of cluster consensus sequences that were used to synthesise a panel of 52,101 Agilent SureSelect RNA probes, each of 120 nucleotides on the complementary strand.

Clinical sample collections for *Castanet* evaluation

We assembled a large collection of samples from two clinical study cohorts in order to assess the performance of our sequencing and analysis workflow.

The GAinS Study (https://ukccggains.com/) is a multi-centre clinical study of sepsis conducted in the UK, including n=2055 sepsis cases for which one or more clinical samples and extensive clinical information were collected [REC: 05/MRE00/38, 08/H0505/78]. We analysed plasma samples from n=573 patients diagnosed with sepsis secondary to community-acquired pneumonia, comprising n=126 'known' cases where clinical microbiology had identified a pathogen (although not necessarily from the available plasma sample) and n=447 'unknown' cases chosen from the whole cohort because no pathogen had been identified in any sample from that case. Accordingly, our sepsis sample collection cannot be considered a random cross-section of cases, nor a full set of samples for each case from which a pathogen might be identified. It was therefore necessary to confirm the pathogen status of the set of sepsis plasma samples used as 'known' samples to validate our methodology. In addition, aggregate results from the available sample collection should not be considered to be representative of the sepsis cohort, with respect to rates of pathogen identification or pathogen frequency.

The ChiMES Study (http://www.encephuk.org/studies/ukchimes) is a multi-centre clinical study of >3,000 children with suspected meningitis and encephalitis conducted in the UK, where one or more clinical samples and extensive clinical information were collected [REC: 11/EM/0442]. We obtained CSF samples from n=243 cases, divided into n=108 'known' cases for which a pathogen had been detected and n=135 'unknown' cases for which no pathogen had been detected in the

Table 1. Organisms included in the *Castanet* probe set.

Virus family	Virus species	Bacterial genus	Bacterial species
Adenoviridae	Human adenovirus	**Acinetobacter**	baumannii
Arenaviridae	Lassa virus		calcoaceticus
	Lymphocytic choriomeningitis virus	**Bartonella**	henselae
Coronaviridae	Human coronavirus HKU1, NL63, OC43, 229E	**Bordetella**	pertussis
	Middle East respiratory syndrome coronavirus	**Borrelia**	burgorferi
	Severe acute respiratory syndrome coronavirus	**Brucella**	spp.
Flaviviridae	Dengue virus	**Burkholderia**	cepacia
	Japanese encephalitis virus	**Chlamydophila**	pneumoniae
	Murray Valley encephalitis virus		psittaci
	St Louis encephalitis virus	**Coxiella**	burnetii
	Tick-borne encephalitis virus	**Enterobacter**	aerogenes
	West Nile virus		cloacae
	Yellow fever virus	**Escherichia**	coli
	Zika virus	**Haemophilus**	influenzae
Herpesviridae	Human herpesvirus 3 (Varicella zoster virus)		parainfluenzae
	Human herpesvirus 4 (Epstein Barr virus)	**Klebsiella**	pneumoniae
	Human herpesvirus 5 (Cytomegalovirus)		oxytoca
	Human herpesvirus 6-7 (Roseolovirus)	**Legionella**	pneumophila
	Human herpesvirus 8 (Kaposi's sarcoma-associated herpesvirus)	**Leptospira**	spp.
	Herpes simplex virus 1-2	**Listeria**	monocytogenes
Orthomyxoviridae	Influenza virus A-C	**Moraxella**	catarrhalis
Paramyxoviridae	Hendra henipavirus	**Mycobacterium**	avium
	Human metapneumovirus		intracellulare
	Human parainfluenzavirus 1-5		tuberculosis
	Measles morbillivirus	**Mycoplasma**	pneumoniae
	Mumps rubulavirus	**Neisseria**	meningitidis
	Nipah henipavirus	**Nocardia**	spp.
	Human respiratory syncytial virus	**Pseudomonas**	aeruginosa
	Sosuga virus	**Serratia**	marcescens
Parvoviridae	Human bocavirus	**Staphylococcus**	aureus
	Human parvovirus B19	**Stenotrophomonas**	maltophilia
	Human parvovirus 4	**Streptococcus**	agalactiae
	Primate erythroparvovirus 1		pneumoniae
	Primate tetraparvovirus 1		pyogenes
Peribunyaviridae	California encephalitis virus	**Treponema**	pallidum
Phenuiviridae	Rift valley fever virus		
	Sandfly fever Naples virus		
	Sandfly fever Sicilian virus		
Picornaviridae	Cardiovirus A-B		
	Coxsackie A virus		
	ECHO virus		
	Enterovirus A, B, D		
	Hepatitis A virus		
	Parechovirus A-B		
	Rhinovirus A-C		
	Rosavirus A		
	Salivivirus		
Polyomaviridae	BK virus		
	JC polyomavirus		
Reoviridae	Rotavirus A-C		
Rhabdovirus	Australian bat lyssavirus		
	Duvenhage lyssavirus		
	European bat lyssavirus 1-2		
	Lagos bat lyssavirus		
	Mokola lyssavirus		
	Rabies lyssavirus		
Togaviridae	Chikungunya virus		
	Eastern equine encephalitis virus		
	Rubella virus		
	Venezuelan equine encephalitis virus		
	Western equine encephalitis virus		

CSF. In general, only one CSF sample had been taken per case, so we were able to assume that our 'known' samples contained the relevant pathogen. Since the complete sample set comprised an essentially random selection from the case cohort, we consider the rates of identification and frequencies of pathogens to be representative of UK childhood meningitis cases within the study period and locations.

Based on the clinical cohort descriptions above, we note that the 'unknown' sample sets are likely to have lower pathogen levels, and/or fewer samples with detectable pathogen material, and/or a higher proportion of pathogens for which effective testing has not been applied. Furthermore, the successful clinical microbiology pathogen identifications among sepsis cases come from a larger number of more diverse sample types (many of which are unavailable for this study) than those in the meningitis cases, emphasising challenges in identifying pathogens among the 'unknown' samples. Further details of cohorts are in the **Supplementary Information**.

A combined library preparation method enables sequencing of both DNA and RNA from microbiological samples

We combined a sample purification method that isolates DNA and RNA together with a single library preparation workflow for both RNA and DNA (**Methods**) in order to minimise sample wastage, reduce costs and avoid bias. We used spike-ins of plasmid DNA and ERCC RNA to confirm that our library method could recover DNA and RNA sequences with similar efficiencies (**Supplementary Information**).

Castanet provides sensitive, quantitative detection and sequencing of viruses and bacteria

To assess the quantitative range of detection of our method, we used dilutions of a commercially available mixture of viruses, designated Viral Multiplex Reference 11/242 (VMR) (National Institute of Biological Standards and Control, UK)[36]. From two undiluted VMR replicates, *Castanet* sequencing yielded $9.1\text{x}10^7$ and $10.9\text{x}10^7$ reads respectively (**Supplementary Information**). All 21 viruses in VMR for which we had enrichment probes were detected, with at least 8.65 reads per million in enriched samples (**Supplementary Figure 1)**. Our method, which has been designed for high-throughput processing of samples in a single assay, compared favourably with the multiple methods adopted for detection of all VMR viruses in 15 other laboratories[36].

We compared viral load with enriched sequence yield for the subset of five VMR viruses that had been quantified by the supplier using qPCR. For each virus, the number of deduplicated *Castanet* reads increased linearly with a similar constant of proportionality to input copy number across a 3-log dilution series (**Figure 1a**). The number of unique reads obtained per virus genome copy differed between viruses, presumably because of differences in sequence target length, the efficiency of sequencing library formation and perhaps, the calibration of qPCR assays.

Similarly, we observed a striking linear relationship between input pathogen load and sequencing yield in a subset of clinical sepsis samples for which we had quantified bacterial and viral load of *Streptococcus pneumoniae* (*Spn*) (n=102) and Epstein-Barr virus (EBV) (n=199) by digital droplet PCR (ddPCR) (**Figure 1b, 1c**). Quantitative yield from targeted enrichment across a wide range of input concentrations is consistent with previous work[28]. In these experiments, sequence enrichment increased the yield of pathogen-specific reads by ~10,000 times for an equivalent amount of sequencing (**Supplementary Figure 2**), implying that batches of a convenient size for larger studies, such as plates of 96 samples prepared and pooled together, would attain at-least-equivalent sensitivity with the same amount of sequencing as a single sample undergoing standard shotgun metagenomic sequencing, massively reducing assay costs.

Large-scale analysis of clinical samples with *Castanet*

We successfully sequenced and included in our analysis a total of 854 samples, derived from 243 ChiMES meningitis cases, 27 non-meningitis negative-control CSF samples, 573 GAinS sepsis cases and 11 negative-control plasma samples. The 243 meningitis cases formed an approximately random sample of disease that comprised 122 for which a pathogen had been identified by clinical microbiology (108 from CSF only plus 14 from a blood sample plus possibly also CSF) and 121 where no pathogen had been identified before sequencing. The sepsis cases comprised 126 for which a pathogen had been identified and 447 chosen for study because no pathogen had been identified in any relevant sample.

Figure 1

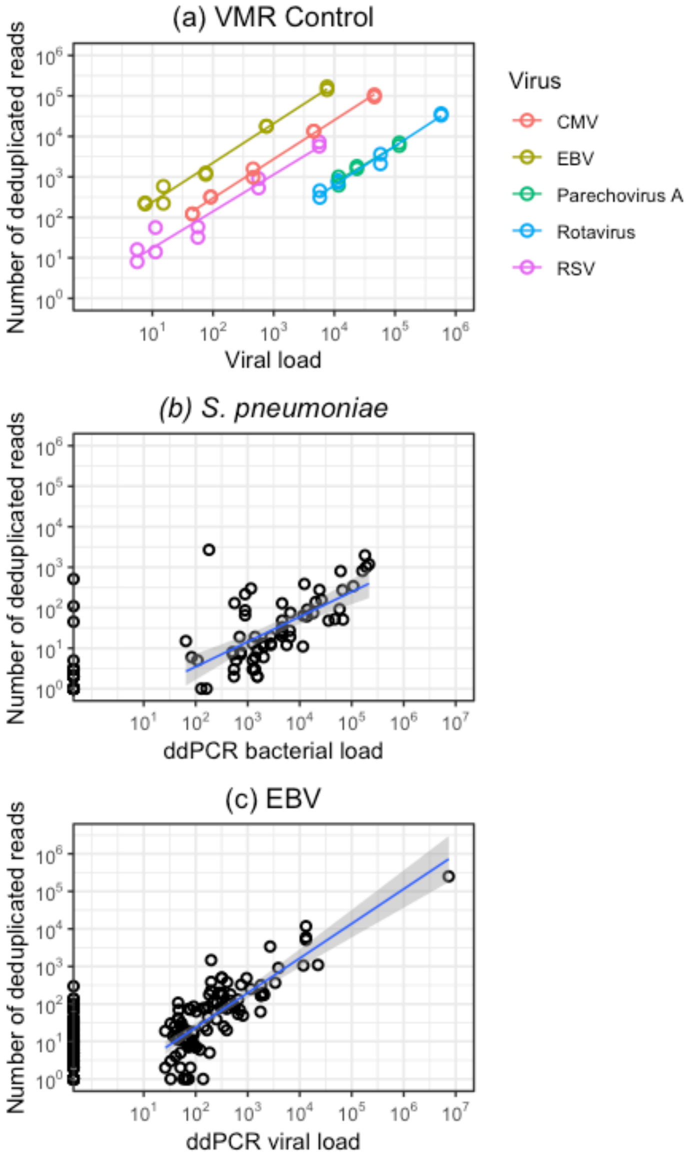


Figure 1: Organism load and sequencing yield in control and sepsis samples. (a) Viral Multiplex Reference reagent was prepared for sequencing at a range of dilutions (neat, 1:10, 1:100, 1:500, 1:1000) in two replicates. Deduplicated read yield was plotted against input viral load for the five viruses that had been quantified by the manufacturer using qPCR. (b) and (c) Pathogen load and sequencing yield in clinical sepsis samples. Deduplicated read yield was plotted against pathogen load, estimated by ddPCR, in samples from a subset of cases in which (b) *Streptococcus pneumoniae* (n=102) or (c) Epstein-Barr Virus (EBV) (n=199) was detected by sequencing. A linear relationship between pathogen load and sequencing yield was observed for each organism (*S. pneumoniae*: $R^2 = 0.449$, $p = 8.8 \times 10^{-5}$; EBV: $R^2 = 0.702$, $p = 2.9 \times 10^{-16}$). The limit of detection of the ddPCR assay was considered to be (b) 46 copies/ml, (c) 26 copies/ml. Sequencing yields for ddPCR-negative samples are shown but not included in the linear regression. VMR, Viral Multiplex Reference; CMV, human cytomegalovirus; EBV, Epstein-Barr virus; RSV, respiratory syncytial virus.

Castanet detects clinically relevant pathogens in CSF and plasma samples

In order to evaluate the ability of *Castanet* to identify pathogens in real clinical samples, we compiled a 'truth dataset' of samples whose status for particular pathogens was known with confidence. For meningitis cases, we accepted the pathogen identification in the case record as the truth state for microbiology-positive samples. For the 126 pathogen-positive CAP sepsis cases, the pathogen identification had often been made in a sample other than plasma and most of the plasma samples in our collection had been obtained after administration of antibiotics, two situations in which plasma levels of pathogens might have been undetectable by any method. Accordingly, we used a positive result by ddPCR for *S. pneumoniae* or EBV to define a sample subset with which to learn the characteristics of pathogen-true-positive sequencing data.

Another key issue in interpreting metagenomic sequencing data, especially in large pools of samples, is to distinguish low-level positives from true-negative samples. We therefore used the opportunity provided by the *S. pneumoniae* ddPCR assays to estimate the threshold for considering a sample sequence-positive in the current study (**Supplementary Figure 3**), deciding to exclude ddPCR-negative samples with fewer than either 72 total or 4 de-duplicated sequence reads from any single pathogen from consideration as sequencing-positive.

Combining the above criteria, we identified 100 ChiMES meningitis CSF and 107 GAinS sepsis plasma samples for inclusion as samples with known pathogen status. To these we added samples included in the cohorts as negative controls (CSF from non-meningitis patients, n=23; and plasma from sepsis-negative patients, n=17) to provide instances of pathogen-negative data. Plasma and CSF samples that were microbiology-positive for a particular pathogen were deemed negative for other pathogens. Reads aligning to viruses known to reactivate in sepsis (herpes simplex virus (HSV), cytomegalovirus (CMV), human herpes virus 6 (HHV6), JC virus) were excluded from the analysis, apart from those EBV samples where ddPCR data was available.

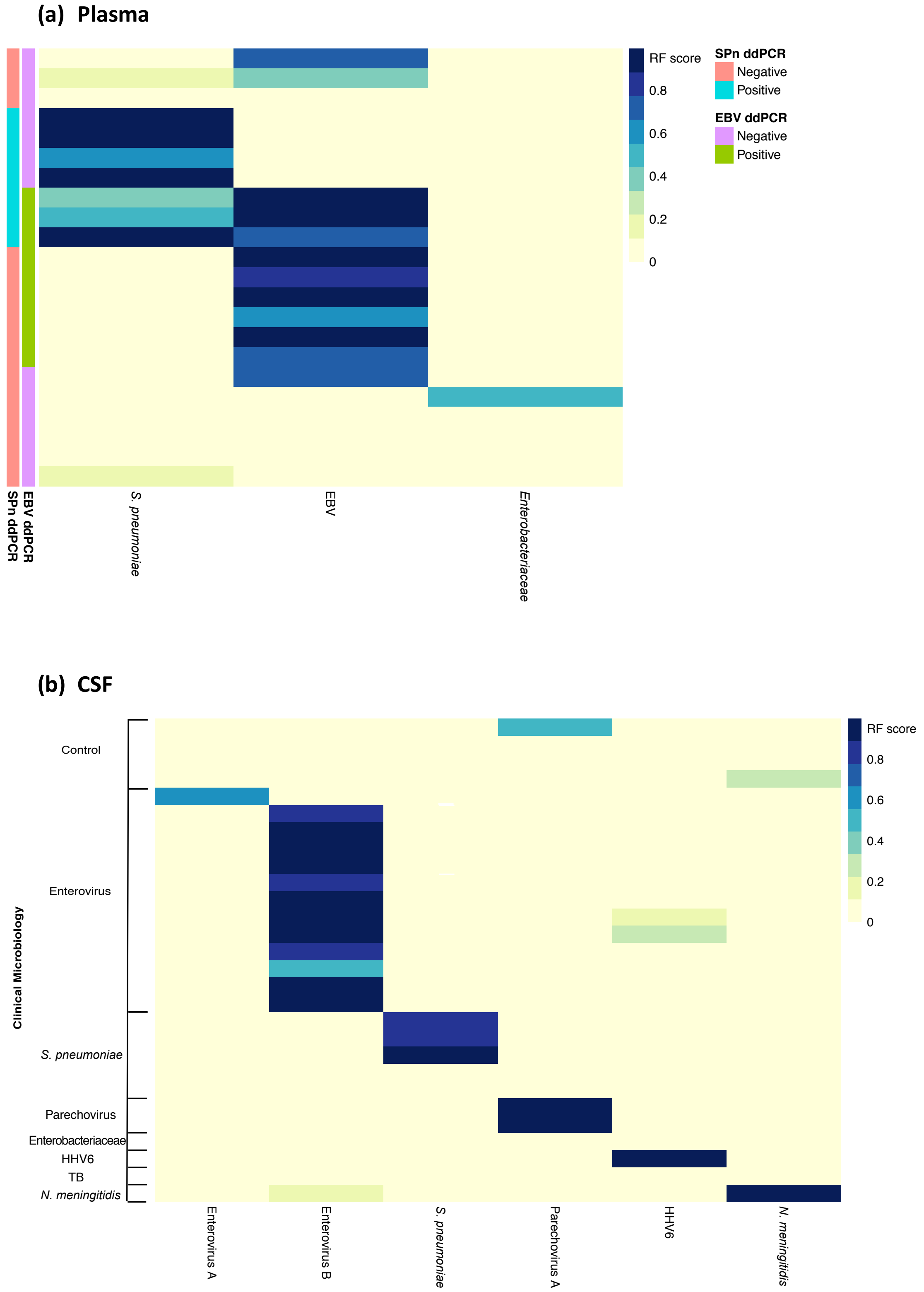
We randomly allocated the 247 samples defined above to training and test datasets in an 80:20 ratio. We used the training dataset (197 samples: 95 CSF; 102 plasma) to train a random forest classifier that used a set of variables derived from the sequencing data (details in **Supplementary Material**) to derive a score between 0 and 1 to indicate whether it was positive for each organism with reads in a sample. Some samples contained reads from multiple organisms and the random forest returns a score for each one of those organisms.

The test dataset comprised 50 samples (28 CSF; 22 plasma). We selected a cut-off random forest score of 0.465 for classification of the test set, to appropriately weight specificity over sensitivity (**Supplementary Figure 4**). At this threshold, there were 5 false negatives and one false positive in the ChiMES test dataset and one false negative and 3 false positives in the GAinS test dataset. In the combined set of test samples the sensitivity was 86.7% (39 of 45 true positives, Figure 2) and the specificity was 98.6% (283 of 287 true negatives). Among the most informative sequencing data metrics for prediction were the numbers of total and deduplicated reads matching a pathogen, taken as the respective proportions of reads aligning to all pathogens in the probeset, and whether a high proportion of the targeted region (regions in our probeset) for that pathogen were covered by reads (**Supplementary Figure 5).** In meningitis predicted-positive samples a median 91% of the genomic target sequence for the implicated pathogen was covered by at least 2 reads. Since excluding pathogens from a diagnosis can also be clinically useful, we assessed the performance of the method in predicting negative status. With a random forest score threshold of 0.015, we predicted 59.2 % of true negatives with a specificity of 97.8%, implying that for many samples it is possible to exclude many possible pathogens without erroneously ruling out true positives.

***Castanet* detects pathogens in samples from patients with no previous pathogen identification**

We analysed the remaining 121 ChiMES meningitis CSF and 447 GAinS sepsis plasma samples, in which no causative pathogen had been identified by routine clinical microbiology. *Castanet* identified one or more pathogens in over one third of samples in both meningitis (41 samples, 34%) and sepsis (165 samples, 37%), including both bacteria and viruses that in many cases were likely to have been causative (**Figure 3**). Among such pathogens, instances of EBV, HHV6, HSV, JC virus and HCMV in sepsis may represent viral reactivation in the context of critical illness, while *Burkholderia cepacia* and *Nocardia asteroides* sequences most probably represent contamination of samples[40]. Excluding these likely reactivations and contaminants, *Castanet* made 39 new detections of clinically relevant pathogens in meningitis patients with negative clinical microbiology, and 50 such new detections in sepsis patients, comprising 32% and 11% of previously unresolved cases respectively (**Table 2** and **Table 3**).

Figure 2


(a) Plasma

(b) CSF

Figure 2: Performance of *Castanet* in clinical samples with a positive microbiology diagnosis. The test dataset included 50 samples: (a) 22 plasma samples (20 GAinS, 2 sepsis-negative controls); (b) 28 CSF samples (24 ChiMES, 4 meningitis-negative controls). The combined overall test dataset specificity and sensitivity was 0.986 and 0.867 respectively. [ChiMES=Childhood Meningitis and Encephalitis Study. GAinS=Genomic Advances in Sepsis. RF=Random Forest. EBV=Epstein-Barr Virus. SPn=Streptococcus pneumoniae. HHV6=human herpes virus 6. ddPCR=droplet digital PCR]. Only organisms detected in each sample set are included as columns in each panel.

Figure 3

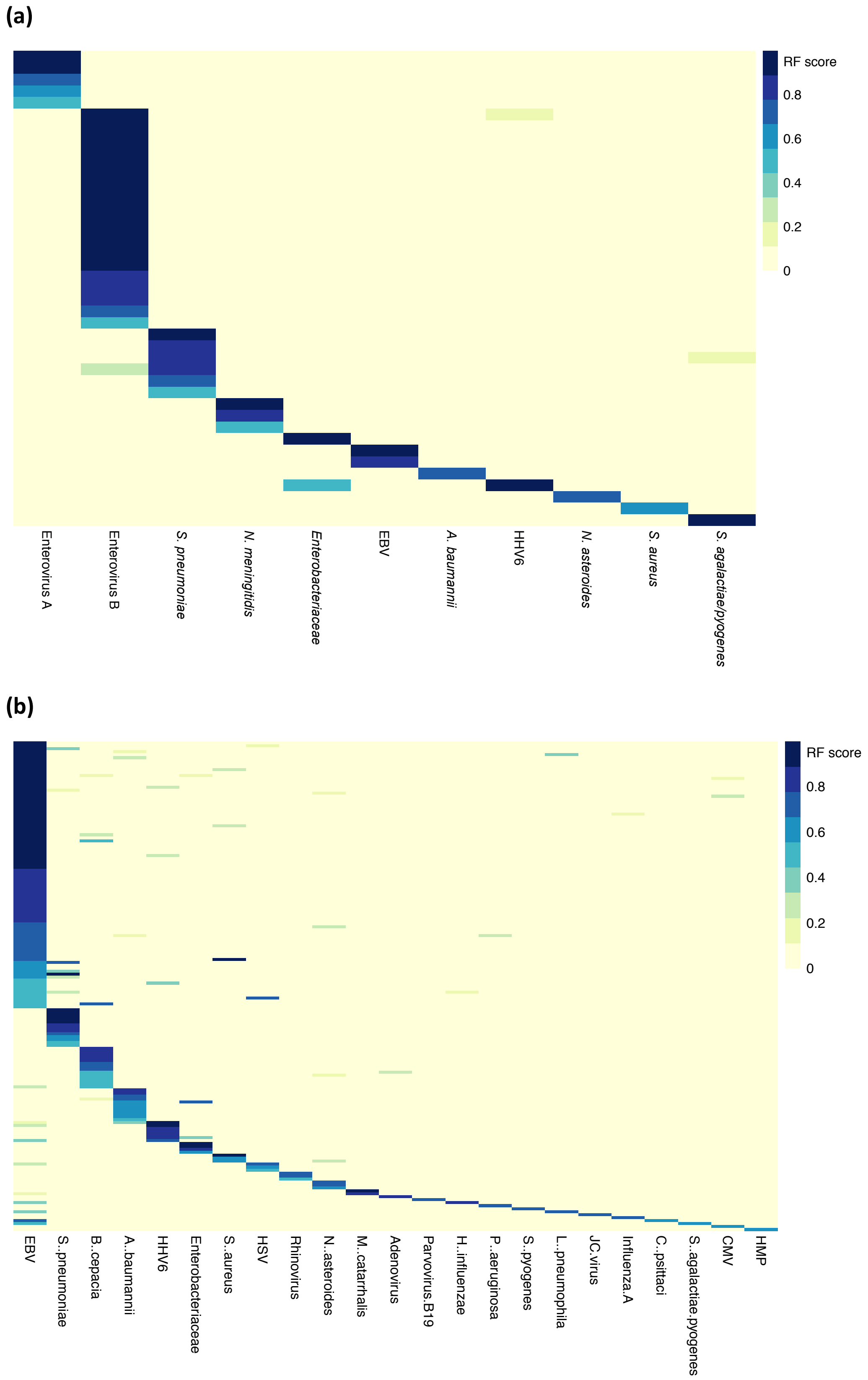


Figure 3: Performance of *Castanet* in samples with no clinical microbiology diagnosis. Panels show as rows (a) 41/121 ChiMES samples; (b) 165/447 GAinS samples; for samples with at least one organism detected at a random forest score RF >0.465. Only organisms detected in each sample set are included as columns in each panel.

Table 2: **New pathogen identifications made by *Castanet* among meningitis CSF samples**

Organism	**Number of patients**
Streptococcus pneumoniae	6
Neisseria meningitidis	3
Enterobacteriaceae	2*
Staphylococcus aureus	1
Streptococcus agalactiae	1
Enterovirus B	19
Enterovirus A	5
EBV	2
HHV6	1*
Total	**39**

Table 3: **New pathogen identifications made by *Castanet* among sepsis plasma samples**

Organism	**Number of patients**
Streptococcus pneumoniae	16
Acinetobacter baumannii	11
Enterobacteriaceae	5
Staphylococcus aureus	4
Moraxella catarrhalis	2
Haemophilus influenzae	1
Pseudomonas aeruginosa	1
Streptococcus pyogenes	1
Legionella pneumophila	1
Chlamydia psittaci	1
Rhinovirus	3
Adenovirus	1
Parvovirus B19	1
Influenza A	1
Human metapneumovirus	1
Total	**50**

In the complete meningitis cohort, *Castanet* improved the overall rate of pathogen identification from 50% (122/243) to 69% (175/243) compared with clinical microbiology alone. Specifically, in CSF samples, *Castanet* identified a likely causative pathogen in 149 of 243 cases (61%) compared with 108/243 (44%) for clinical microbiology (**Table 4**). **Table 5** gives equivalent totals for the sepsis plasma sample set.

Table 4. Pathogens identified in all meningitis clinical cases / meningitis CSF samples.

Organism	**Clinical microbiology: CSF only**	**Clinical microbiology: CSF + blood**	**Clinical microbiology + *Castanet***	**Clinical microbiology + *Castanet***	***Castanet***	**Prevalence in sample set**
Unknown	135 (56%)	121 (50%)	80	80 (33%)	94 (37%)	42%
Enterovirus	45 (19%)	45 (19%)	75	163 (67%)	74 (30%)	31%
Human parechovirus	14 (6%)	14 (6%)	15		13 (5%)	4%
Streptococcus pneumoniae	11 (5%)	15 (6%)	24		18 (7%)	5%
Neisseria meningitidis	8 (3%)	13 (5%)	17		15 (6%)	6%
Other	30 (14%)	35 (15%)	44		29 (12%)	11%
Total	**243**	**243**	**255 ***	**243**	**243**	

* 12 samples called differently by lab and *Castanet* appear in two categories in the column above.

Table 5. Pathogens identified in all sepsis clinical cases / sepsis plasma samples.

Organism	**Clinical microbiology**	**Clinical microbiology + *Castanet***	***Castanet***	**Prevalence in sample set**
Unknown	447 (78%)	397 (69%)	488 (85%)	60%
Streptococcus pneumoniae	92 (16%)	108 (19%)	45 (8%)	15%
Other bacterial	6 (1%)	33 (6%)	31 (5%)	18%
Influenza	13 (2%)	14 (2%)	1 (0%)	5%
Other viral	15 (3%)	21 (4%)	8 (1%)	3%
Total	**573**	**573**	**573**	

In several cases sequencing suggested results that differed from the recorded clinical microbiology. In some instances, sample-specific reads corresponding to the microbiological diagnosis remained below the threshold of the detection model. Many discrepancies were difficult to resolve (for example where a different pathogen may have actually been present in another sample from the same case), so we focused on Enterovirus detections, where there was good information on PCR-based testing, at least at the case level (**Table 6**). There are 18 samples which are PCR-negative, but *Castanet*-positive.

Table 6. Resolving Enterovirus identifications in the ChiMES meningitis CSF sample set

	Clinical microbiology status			
	PCR-positive	PCR-negative	Not tested	**Total**
Castanet-positive	44	18	16	**78**
Castanet-negative	1	142	50	**193**
Total	**45**	**160**	**66**	**271**

Our investigation provided a simple example of one potential application of the methodology: more extensive, comprehensive testing for a wider panel of potential pathogens can lead us to shift our understanding of (i) the complement and (ii) the frequency distribution of pathogens causing a disease in a specific population. In the meningitis cohort we detected (and in most cases derived a complete genome sequence for) a substantial number of diverse enteroviruses, such that the estimated frequency of enteroviruses among all identified pathogens (74 Enterovirus, 75 other pathogens) in the sample defined using *Castanet* was significantly higher than that defined using clinical microbiology methods alone (45 Enterovirus, 77 other pathogens; p = 0.03, Fisher's Exact Test).

Discussion

Our workflow, *Castanet*, is a versatile probe-based enrichment sequencing method that addresses several of the issues associated with sequencing for pathogen detection. Our method is sensitive and cost-effective, combining the analysis of RNA and DNA from the same material in a single protocol and enriching for pathogens of interest using a modestly sized panel of probes that efficiently targets a diverse range of viral and bacterial pathogens and can be tailored to particular settings.

Castanet attains molecular sensitivity comparable to PCR with a yield of pathogen sequence reads that is proportional to pathogen load. In this study we have adopted a probe panel targeting members of 17 virus families and 35 bacterial species relevant to two disease presentations. With approximately 10,000-times enrichment of targeted sequences, the method can analyse many samples in parallel using the same amount of sequencing that could otherwise be needed for one sample.

We evaluated the *Castanet* methodology on a large sample set from two densely phenotyped disease cohorts collected across multiple centres in the UK. We selected these cohorts as the most challenging situations on which an assay like *Castanet* would be applied in a diagnostic laboratory setting. In both sepsis and meningitis, the absolute rate of identification of pathogens tends to be low despite extensive laboratory testing, and many samples may be taken after administration of antibiotics. In the setting of sepsis secondary to pneumonia, plasma may also contain little or no pathogen material from agents such as influenza virus A that had been responsible for the initial infection.

In spite of difficulties in describing a ground-truth dataset of known positives and known negatives, we clearly demonstrate that *Castanet* embodies high sensitivity and specificity in a single test. Acknowledging the complexities of interpreting metagenomics data, we show that the characteristics of pathogen-positive and pathogen-negative samples can be defined using a random forest model that combines data across pathogens and diseases.

A key aim of our study was to evaluate the performance of the *Castanet* method in increasing the number of pathogen identifications made from challenging samples. We identify a substantial number of previously unrecognised infections and show that *Castanet* can be used to update estimates of population pathogen frequencies derived from conventional testing.

The *Castanet* method has a variety of prospective applications. A key difference from other targeted assays is that there is no need to decide for each sample what tests to perform; with probe-enrichment the user can in principle (and in practice) construct a very long list of plausible pathogens to test for with equivalent sensitivity. By providing a wide specificity in a single test on large batches of samples, *Castanet* will be useful in investigating the performance and coverage of existing testing regimens and platforms (for example to check whether a test covers all circulating variants or the set of tests in use includes all common pathogens in the population). Similarly, *Castanet* could define with minimal bias the distribution of pathogens associated with a disease in a particular population, particularly for low- and middle-income countries where treatment may be based on untested assumptions. In principle, probe enrichment sequencing could allow clinicians for the first time to exclude many possible causes of infection at the same time, in contrast to current methodologies.

In this study, we provide an example of a further use of the *Castanet* workflow by deriving complete genomes of diverse viruses on a large scale, with sensitivity comparable to PCR but with greater robustness and without the requirement for assay optimisation. In addition to the diverse set of Enterovirus genomes generated within this study we have started using *Castanet* as a high-throughput pipeline for sequencing viruses from nasal swab samples in acute respiratory infection (manuscript in preparation).

Future work is envisaged to achieve better resolution of pathogenic species within clusters of closely related bacteria such as the Gram-negatives, Streptococci and Staphylococci, including by focusing on bacterial sequences of particular clinical interest such as factors associated with virulence and anti-microbial resistance. Further challenges posed by the need to optimise sensitivity by improving the efficiency with which pathogen nucleic acids are processed into sequencing libraries that could also incorporate unique molecular identifiers to enhance subsequent data analysis.

Metagenomics using short read sequencing has long promised a new approach to microbiological testing, with limited impact as a patient-focused test because of unwieldy preparation methods and slow, costly and high-volume sequencing runs. On the other hand, streamlined sample-processing and immediate dataflows mean that nanopore-based sequencing is already close to providing a workflow compatible with a fast-turnaround lab-based or point-of-care test, even if data volumes are not yet sufficient for sensitive metagenomics at a reasonable cost. Probe enrichment can provide benefits for both approaches, enabling high-throughput batch testing and, if capture times can be reduced, enhancing the sensitivity and therefore speed and cost of nanopore sequencing. In either form, our approach could be useful in routine microbiology as a backup to fast-turnaround tests or to exclude large numbers of rare pathogens with turnaround of results in a few days.

Castanet and related targeted metagenomics methodologies provide a significant step towards the widespread use of metagenomics in identifying and characterising pathogen sequences in large-scale clinical studies. By providing a single workflow, *Castanet* promises to simplify and generally enable phylo-epidemiological studies on many viruses and to be valuable in defining causes of infection at the population level, with benefits for the quality of microbiological testing.

Methods

Experimental procedures

Nucleic acid extraction Total nucleic acids were extracted using the NucliSENS easyMag platform (Biomerieux) from up to 500ul of plasma or CSF where available and the extracted nucleic acids were eluted in 25ul of kit elution buffer and stored at -80°C.

Library preparation RNA was reverse-transcribed with random priming using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs) with 5ul sample input. We made modifications to the manufacturer's guidelines, including fragmentation for 4 minutes at 94°C and omission of Actinomycin D at first-strand reverse transcription, and we stopped the protocol after generation of double-stranded cDNA. The resulting mixture of cDNA and sample DNA (normalised to 5ng/ul or the maximum available) was processed using the Nextera DNA Library Preparation Kit (Illumina) according to the manufacturer's guidelines, including library amplification for 15 PCR cycles using custom indexed primers and post-PCR clean-up with 0.85x volume Ampure XP (Beckman Coulter). Libraries were quantified using Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen) and analysed using Agilent TapeStation with D1K High Sensitivity Kit (Agilent) for equimolar pooling.

Probe-based enrichment 1ug of each indexed pooled library was enriched using the Agilent SureSelect^{XT} Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library protocol with one major modification to the recommended protocol: the capture was performed on the post-PCR indexed pool, using oligonucleotide blockers complementary to adapter sequences.

Sequencing Sequencing was performed on the Illumina MiSeq or HiSeq 4000 platforms generating either 75-bp or 150-bp paired-end reads.

Digital Droplet PCR (ddPCR) Aliquots of extracted nucleic acid from plasma were processed in triplicate (1.5ul per replicate) following the recommended workflow (WX200 ddPCR system, Bio-Rad). Custom PrimeTime

(IDT) primer/probe sets targeting *S.* pneumoniae,[41] and Epstein-Barr Virus (EBV),[42] were designed based on published sequence data. Full details are included in the Supplementary Material.

Bioinformatics and statistical analysis

Sepsis and meningitis samples and uninfected controls: De-multiplexed sequence read-pairs were trimmed of adapter sequences using Trimmomatic v0.36, with the ILLUMINACLIP options set to "2:10:7:1:true MINLEN:80", using the set of Illumina adapters supplied with the software.[43] The trimmed reads were then classified using Kraken v1[44] using a custom database containing the human genome (GRCh38 build), all RefSeq bacterial and viral genomes, and a selection of fungal genomes that were most likely to be associated with cases of meningitis[45]. These were: *Aspergillus fumigatus*, *Candida* spp., Coccidioides spp., *Cryptococcus* spp., *Histoplasma capsulatum*, *Paracoccidioides brasiliensis*, and *Pneumocystis* spp. Reads identified as bacterial or viral and unclassified reads were aligned using BWA v0.7.12[46] with default settings to a multi-fasta reference of consensus sequences corresponding to the enrichment probe targets, augmented with sequences of known or suspected contaminants. These included (i) reagent contaminants (Alteromonas and Achromobacter spp.), (ii) genomes of two viruses known to have been sequenced on the same flow cell: MVMPCG and Echovirus 7; and (iii) the rMLST sequences of commensal *Streptococcus* species species that were thought to be likely contaminants in clinical samples (Supplementary data file).

Duplicate reassignment: We first corrected our sequencing results for index misassignment, a well-recognised issue with multiplexed sequencing where a small proportion of reads in each sample represents an incorrectly identified read that had come from a nearby optical cluster. For each sequencing pool, we identified PCR-duplicated reads and reassigned all reads in each duplicate cluster to the sample with the highest number of reads in that cluster.

Collection of sequencing statistics: Following duplicate reassignment, for each sample and target organism we calculated a set of descriptive statistics, which included sequencing depth with and without deduplication, and coverage of target sequences at a various depth thresholds. The collected statistics were combined with available laboratory data on sample positivity, and with ddPCR results where these were available. The resulting data frame was used to train a Random Forest model.

Random Forest classification model: The function randomForest from the R package of the same name was used to derive an algorithm to classify reads aligning to each organism in a sample as positive or negative (details in **Supplementary Information**).

References

1. Dellinger RP, Levy MM, Rhodes A, et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. *Crit Care Med* 2013;41(2):580-637. doi: 10.1097/CCM.0b013e31827e83af [published Online First: 2013/01/29]
2. Gudina EK, Tesfaye M, Wieser A, et al. Outcome of patients with acute bacterial meningitis in a teaching hospital in Ethiopia: A prospective study. *PLoS ONE* 2018;13(7):e0200067. doi: 10.1371/journal.pone.0200067 [published Online First: 2018/07/19]
3. McGill F, Heyderman RS, Panagiotou S, et al. Acute bacterial meningitis in adults. *Lancet* 2016;388(10063):3036-47. doi: 10.1016/s0140-6736(16)30654-7 [published Online First: 2016/06/07]
4. Scarborough M, Thwaites GE. The diagnosis and management of acute bacterial meningitis in resource-poor settings. *Lancet Neurol* 2008;7(7):637-48. doi: 10.1016/s1474-4422(08)70139-x [published Online First: 2008/06/21]
5. Tadesse BT, Foster BA, Shibeshi MS, et al. Empiric Treatment of Acute Meningitis Syndrome in a Resource-Limited Setting: Clinical Outcomes and Predictors of Survival or Death. *Ethiopian*

journal of health sciences 2017;27(6):581-88. doi: 10.4314/ejhs.v27i6.3 [published Online First: 2018/03/01]

6. Ripa T, Nilsson P. A variant of Chlamydia trachomatis with deletion in cryptic plasmid: implications for use of PCR diagnostic tests. *Euro Surveill* 2006;11(11):E061109.2. doi: 10.2807/esw.11.45.03076-en [published Online First: 2007/01/11]
7. Xu M, Qin X, Astion ML, et al. Implementation of filmarray respiratory viral panel in a core laboratory improves testing turnaround time and patient care. *American journal of clinical pathology* 2013;139(1):118-23. doi: 10.1309/ajcph7x3nlyzphbw [published Online First: 2012/12/29]
8. Lee SH, Chen SY, Chien JY, et al. Usefulness of the FilmArray meningitis/encephalitis (M/E) panel for the diagnosis of infectious meningitis and encephalitis in Taiwan. *J Microbiol Immunol Infect* 2019 doi: 10.1016/j.jmii.2019.04.005 [published Online First: 2019/05/16]
9. Gadsby NJ, Russell CD, McHugh MP, et al. Comprehensive Molecular Testing for Respiratory Pathogens in Community-Acquired Pneumonia. *Clin Infect Dis* 2016;62(7):817-23. doi: 10.1093/cid/civ1214 [published Online First: 2016/01/10]
10. Leber AL, Everhart K, Balada-Llasat JM, et al. Multicenter Evaluation of BioFire FilmArray Meningitis/Encephalitis Panel for Detection of Bacteria, Viruses, and Yeast in Cerebrospinal Fluid Specimens. *J Clin Microbiol* 2016;54(9):2251-61. doi: 10.1128/jcm.00730-16 [published Online First: 2016/06/24]
11. Tokarz R, Mishra N, Tagliafierro T, et al. A multiplex serologic platform for diagnosis of tick-borne diseases. *Sci Rep* 2018;8(1):3158. doi: 10.1038/s41598-018-21349-2 [published Online First: 2018/02/18]
12. Brenner N, Mentzer AJ, Butt J, et al. Validation of Multiplex Serology detecting human herpesviruses 1-5. *PLoS ONE* 2018;13(12):e0209379. doi: 10.1371/journal.pone.0209379 [published Online First: 2018/12/28]
13. Wellinghausen N, Kochem AJ, Disque C, et al. Diagnosis of bacteremia in whole-blood samples by use of a commercial universal 16S rRNA gene-based PCR and sequence analysis. *J Clin Microbiol* 2009;47(9):2759-65. doi: 10.1128/jcm.00567-09 [published Online First: 2009/07/03]
14. Decuypere S, Meehan CJ, Van Puyvelde S, et al. Diagnosis of Bacterial Bloodstream Infections: A 16S Metagenomics Approach. *PLoS Negl Trop Dis* 2016;10(2):e0004470. doi: 10.1371/journal.pntd.0004470 [published Online First: 2016/03/02]
15. Wilson MR, Sample HA, Zorn KC, et al. Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. *N Engl J Med* 2019;380(24):2327-40. doi: 10.1056/NEJMoa1803396 [published Online First: 2019/06/13]
16. Grumaz S, Stevens P, Grumaz C, et al. Next-generation sequencing diagnostics of bacteremia in septic patients. *Genome Med* 2016;8(1):73. doi: 10.1186/s13073-016-0326-8 [published Online First: 2016/07/03]
17. Sharp C, Golubchik T, Gregory WF, et al. 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. *BMC Res Notes* 2018;11(1):120. doi: 10.1186/s13104-018-3234-8 [published Online First: 2018/02/11]
18. Ivy MI, Thoendel MJ, Jeraldo PR, et al. Direct Detection and Identification of Prosthetic Joint Infection Pathogens in Synovial Fluid by Metagenomic Shotgun Sequencing. *J Clin Microbiol* 2018;56(9) doi: 10.1128/jcm.00402-18 [published Online First: 2018/06/01]

19. Naccache SN, Greninger AL, Lee D, et al. The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J Virol* 2013;87(22):11966-77. doi: 10.1128/jvi.02323-13 [published Online First: 2013/09/13]
20. Allander T, Emerson SU, Engle RE, et al. A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc Natl Acad Sci U S A* 2001;98(20):11609-14. doi: 10.1073/pnas.211424698 [published Online First: 2001/09/20]
21. Charalampous T, Kay GL, Richardson H, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat Biotechnol* 2019;37(7):783-92. doi: 10.1038/s41587-019-0156-5 [published Online First: 2019/06/27]
22. Duhaime MB, Sullivan MB. Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* 2012;434(2):181-6. doi: 10.1016/j.virol.2012.09.036 [published Online First: 2012/10/23]
23. Breitbart M, Rohwer F. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *Biotechniques* 2005;39(5):729-36. doi: 10.2144/000112019 [published Online First: 2005/11/30]
24. Okou DT, Steinberg KM, Middle C, et al. Microarray-based genomic selection for high-throughput resequencing. *Nat Methods* 2007;4(11):907-9. doi: 10.1038/nmeth1109 [published Online First: 2007/10/16]
25. Lovett M, Kere J, Hinton LM. Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc Natl Acad Sci U S A* 1991;88(21):9628-32. [published Online First: 1991/11/01]
26. Albert TJ, Molla MN, Muzny DM, et al. Direct selection of human genomic loci by microarray hybridization. *Nat Methods* 2007;4(11):903-5. doi: 10.1038/nmeth1111 [published Online First: 2007/10/16]
27. Hodges E, Xuan Z, Balija V, et al. Genome-wide in situ exon capture for selective resequencing. *Nat Genet* 2007;39(12):1522-7. doi: 10.1038/ng.2007.42 [published Online First: 2007/11/06]
28. Bonsall D, Ansari MA, Ip C, et al. ve-SEQ: Robust, unbiased enrichment for streamlined detection and whole-genome sequencing of HCV and other highly diverse pathogens. *F1000Res* 2015;4:1062. doi: 10.12688/f1000research.7111.1 [published Online First: 2015/01/01]
29. Depledge DP, Palser AL, Watson SJ, et al. Specific capture and whole-genome sequencing of viruses from clinical samples. *PloS one* 2011;6(11):e27805. doi: 10.1371/journal.pone.0027805 [published Online First: 2011/11/30]
30. Duncavage EJ, Magrini V, Becker N, et al. Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. *J Mol Diagn* 2011;13(3):325-33. doi: 10.1016/j.jmoldx.2011.01.006 [published Online First: 2011/04/19]
31. Koehler JW, Hall AT, Rolfe PA, et al. Development and evaluation of a panel of filovirus sequence capture probes for pathogen detection by next-generation sequencing. *PLoS ONE* 2014;9(9):e107007. doi: 10.1371/journal.pone.0107007 [published Online First: 2014/09/11]
32. Briese T, Kapoor A, Mishra N, et al. Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. *mBio* 2015;6(5)
33. Brown AC, Bryant JM, Einer-Jensen K, et al. Rapid Whole Genome Sequencing of M. tuberculosis directly from clinical samples. *J Clin Microbiol* 2015
34. Allicock OM, Guo C, Uhlemann A-C, et al. BacCapSeq: a Platform for Diagnosis and Characterization of Bacterial Infections. *mBio* 2018;9(5):e02007-18. doi: 10.1128/mBio.02007-18

35. Metsky HC, Siddle KJ, Gladden-Young A, et al. Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. *Nat Biotechnol* 2019;37(2):160-68. doi: 10.1038/s41587-018-0006-x [published Online First: 2019/02/06]

36. Mee ET, Preston MD, Minor PD, et al. Development of a candidate reference material for adventitious virus detection in vaccine and biologicals manufacturing by deep sequencing. *Vaccine* 2016;34(17):2035-43. doi: 10.1016/j.vaccine.2015.12.020

37. Gupta S, Sakhuja A, Kumar G, et al. Culture-Negative Severe Sepsis: Nationwide Trends and Outcomes. *Chest* 2016;150(6):1251-59. doi: 10.1016/j.chest.2016.08.1460 [published Online First: 2016/09/13]

38. Katoh K, Misawa K, Kuma K, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;30(14):3059-66. [published Online First: 2002/07/24]

39. Jolley KA, Bliss CM, Bennett JS, et al. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* 2012;158(Pt 4):1005-15. doi: 10.1099/mic.0.055459-0 [published Online First: 2012/01/28]

40. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology* 2014;12:87. doi: 10.1186/s12915-014-0087-z [published Online First: 2014/11/13]

41. Park HK, Lee HJ, Kim W. Real-time PCR assays for the detection and quantification of Streptococcus pneumoniae. *FEMS Microbiol Lett* 2010;310(1):48-53. doi: 10.1111/j.1574-6968.2010.02044.x [published Online First: 2010/07/17]

42. Ryan JL, Fan H, Glaser SL, et al. Epstein-Barr virus quantitation by real-time PCR targeting multiple gene segments: a novel approach to screen for the virus in paraffin-embedded tissue and plasma. *J Mol Diagn* 2004;6(4):378-85. doi: 10.1016/s1525-1578(10)60535-1 [published Online First: 2004/10/28]

43. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30(15):2114-20. doi: 10.1093/bioinformatics/btu170 [published Online First: 2014/04/04]

44. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;15(3):R46. doi: 10.1186/gb-2014-15-3-r46 [published Online First: 2014/03/04]

45. Cuomo CA. Harnessing Whole Genome Sequencing in Medical Mycology. *Current fungal infection reports* 2017;11(2):52-59. doi: 10.1007/s12281-017-0276-7 [published Online First: 2017/09/15]

46. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25(14):1754-60. doi: 10.1093/bioinformatics/btp324 [published Online First: 2009/05/20]

Study Information

Contributions

C.G., T.G., A.A., A.T., D.B., P.P., E.B., A.P., M.S., J.K., and R.B. conceived and designed the study. C.H. and P.H. (GAinS), and M.S., N.M., S.D., M.G., A.P., and T.S. (ChiMES) were responsible for clinical study setup, patient recruitment and sample collection. A.A. designed the probe set. C.G., I.E., J.B., M.M., M.S. and R.B. contributed to probe design. C.G., M.C., A.T., D.B., P.P., and R.B. contributed to laboratory methods development. A.B. and C.G. undertook the nucleic acid extractions. M.C., A.T., H.S., and C.G. undertook the library preparations. C.G., T.G., R.B. and A.A. analysed sequence data.

T.G. designed and developed the computational tools. J.K., A.P. and M.S. contributed to data analysis and interpretation. C.G., T.G., A.A., and R.B. wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

We thank Annabel Coxon, Gretchen Meddaugh and Rebecca Beckley for research support.

Research Ethics

ChiMES: part of the larger ENCEPH UK study - NRES Committee East Midlands - Nottingham 1 - 11/EM/0442.

GAinS: REC: 05/MRE00/38 (Scotland A Research Ethics Committee), 08/H0505/78 (Berkshire Research Ethics Committee).

Funding

This work was supported by a grant from Meningitis Research UK and by the Wellcome Trust via the Oxford-Wellcome Institutional Strategic Support Fund and core funding to the Wellcome Centre for Human Genetics (awards 090532/Z/09/Z and 203141/Z/16/Z). C.G. was supported by a Medical Research Council/British Journal of Anaesthesia/Royal College of Anaesthetists Fellowship (MR/M018725/1). M.A. was funded by the Oxford Martin School for part of this project. J.K. was supported by a Wellcome Trust Investigator Award (204969/Z/16/Z) and the NIHR Oxford Biomedical Research Centre. E.B. was funded by the Medical Research Council UK and the Oxford NIHR Biomedical Research Centre and is an NIHR Senior Investigator. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. Financial support was provided by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z.

The views expressed in this report are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.