scientific reports

Check for updates

OPEN

Optimization of cerebrospinal fluid microbial DNA metagenomic sequencing diagnostics

Josefin Olausson[1,2], Sofia Brunet[1], Diana Vracar[1,2], Yarong Tian[2], Sanna Abrahamsson[2], Sri Harsha Meghadri[2], Per Sikora[3], Maria Lind Karlberg[4], Hedvig E. Jakobsson[1,2]✉ & Ka-Wei Tang[1,2]

Infection in the central nervous system is a severe condition associated with high morbidity and mortality. Despite ample testing, the majority of encephalitis and meningitis cases remain undiagnosed. Metagenomic sequencing of cerebrospinal fluid has emerged as an unbiased approach to identify rare microbes and novel pathogens. However, several major hurdles remain, including establishment of individual limits of detection, removal of false positives and implementation of universal controls. Twenty-one cerebrospinal fluid samples, in which a known pathogen had been positively identified by available clinical techniques, were subjected to metagenomic DNA sequencing. Fourteen samples contained minute levels of Epstein-Barr virus. The detection threshold for each sample was calculated by using the total leukocyte content in the sample and environmental contaminants found in the bioinformatic classifiers. Virus sequences were detected in all ten samples, in which more than one read was expected according to the calculations. Conversely, no viral reads were detected in seven out of eight samples, in which less than one read was expected according to the calculations. False positive pathogens of computational or environmental origin were readily identified, by using a commonly available cell control. For bacteria, additional filters including a comparison between classifiers removed the remaining false positives and alleviated pathogen identification. Here we show a generalizable method for identification of pathogen species using DNA metagenomic sequencing. The choice of bioinformatic method mainly affected the efficiency of pathogen identification, but not the sensitivity of detection. Identification of pathogens requires multiple filtering steps including read distribution, sequence diversity and complementary verification of pathogen reads.

Infections in the central nervous system (CNS) are severe and despite extensive microbiological diagnostic analysis a causative pathogen cannot be identified in many of the cases. A majority of CNS infections are caused by viruses, such as herpes simplex virus 1 (HSV1), varicella zoster virus (VZV or human herpesvirus 3) and enterovirus[1,2]. Among bacterial CNS infections, *Streptococcus pneumoniae* and *Neisseria meningitidis* are the most common pathogens, while fungal or parasitic meningitis CNS infections are less common[3]. Epstein-Barr virus (EBV) has been implicated in recurrent meningitis and chronic encephalitis[4]. However, due to the high prevalence of EBV and its ability to remain latent in B-lymphocytes after primary infection and its role in tumorigenesis, assessing the clinical relevance of EBV DNA detected in cerebrospinal fluid (CSF) is difficult and presence of EBV is often considered to be a benign incidental finding[5,6].

Current microbiological diagnostic methods include cultivation and nucleic acid detection of CSF, which are restricted to prior knowledge of the putative causing agent. Cultivation can detect a wide range of microorganisms; however, it is limited to viable and culturable pathogens. In contrast, nucleic acid detection is rapid and highly sensitive, but constrained to genetically conserved regions of known pathogens. Metagenomic sequencing using massive parallel sequencing, has the capability to discern multiple species and identify unknown species in samples. In metagenomics, the total nucleic acid present in the clinical sample is sequenced, thus providing an unbiased tool to diagnose infections and unknown species in samples[7–9].

[1]Department of Clinical Microbiology, Region Västra Götaland, Sahlgrenska University Hospital, Gothenburg, Sweden. [2]Department of Infectious Diseases, Wallenberg Centre for Molecular and Translational Medicine, Institute of Biomedicine, University of Gothenburg, Gothenburg, Sweden. [3]Clinical Genomics Gothenburg, Science for Life Laboratories, Gothenburg, Sweden. [4]Department of Microbiology, Public Health Agency of Sweden, Solna, Sweden. ✉email: hedvig.jakobsson@gu.se

While there currently is no universal standard for metagenomic sequencing in a clinical setting and the technique is still faced with some major challenges[10], there are studies using different approaches attempting to standardize the analysis[11–13]. In addition, metagenomic sequencing guidelines have recently been published for viral metagenomics in a clinical setting[14–16].

Contrary to PCR, the sensitivity in metagenomic sequencing is dependent on the fraction of pathogen sequences in the total sequencing library. Furthermore, laboratory contaminants detected in sequencing have been shown to differ greatly between laboratories[17]. Nucleic acid derived from the host and environmental contaminants must therefore be taken into account. Previous studies have calculated the sensitivity by using dilution of an exogenous pathogen into a known quantity of host background. While this gives an estimation of sensitivity, it does not take into account the variability of total nucleic acid content in clinical samples nor does it provide any guidance on how the sensitivity of each sample might be calculated prior to sequencing. However, spike-in controls are a well-tested approach to conclude reproducibility and separate contaminants from true pathogens[13,18].

Bioinformatic pathogen identification is a second major obstacle. Several publicly available bioinformatic tools for classification are available, such as Centrifuge, Kraken, PAIPline and PathSeq[19–22]. Two conceptually different methods are frequently used, alignment of single reads (e.g. BLAST), or assemblies (k-mers), against pathogen databases. The list of pathogens generated by these applications are often long and require exhaustive examinations in order to discern the true pathogen from bioinformatic misclassification and environmental contaminants. Criteria for identifying the causative pathogen include sequences disseminated throughout the microbial genome of the proposed pathogen, a threshold for number of pathogen reads in relation to total number of reads, and confirmation using several alignment algorithms have been suggested to increase the specificity[23,24]. Each laboratory does however apply their own criteria.

We investigated the robustness of microbial DNA metagenomics for clinical diagnostics of CNS-infections. To evaluate the diagnostic performance of the method, 21 CSF samples with known pathogens levels originating from patients with variable levels of leukocyte content were sequenced with the aim to identify factors important for calculating sensitivity. Also, four different taxonomic classifiers were assessed for their efficiency to identify pathogens as well as the number of false positive pathogens identified. Two commonly available cell lines were implemented as a positive and negative control to support the removal of environmental contaminants and bioinformatic misclassifications.

Results

We implemented a metagenomic DNA sequencing methodology to unbiasedly detect microbial species in CSF samples from patients with CNS symptoms in which a pathogen or EBV had been detected (Additional 3: Table 1). Samples positively identified with pathogen-specific quantitative PCR (qPCR), 16S rRNA gene sequencing or bacterial/mycotic culture in CSF were included. Different pathogen types and variation of viral loads were chosen. The majority of samples CSF contained low levels of EBV, but variable levels of leukocyte content. These samples were chosen to establish the sensitivity of the method. DNA from each sample was extracted and fragmented before library preparation and sequenced using massive parallel sequencing. An overview of the experimental procedure, bioinformatic classifiers, and post-classification analyses are depicted in Fig. 1.

Bioinformatic classifiers. Four bioinformatic classifiers were included, Kraken2, Centrifuge, our in-house developed PaRCA (Pathogen detection for Research and Clinical Applications) and CosmosID. CosmosID was tested mainly for its ability to generate concise pathogen lists, but the format of the platform prevented a detailed analysis of the raw data and was therefore not included in all comparisons in the manuscript. The four bioinformatic classifiers diverged with regards to fraction of processed reads (from 85 to 100%, Additional 3: Tables 2–3). However, the ability to identify the primary pathogen was similar when comparing the classifiers.

Sensitivity. Initially, three CSF samples (Sample 1–3) with high virus load of herpesvirus were analyzed. HSV1 and VZV were detected by all bioinformatic classifiers (Table 1). In sample 1, HSV1 was positively identified at 1×10^4 genome equivalents per milliliter (Geq/ml) using qPCR. The sequencing library consisting of more than 15 million reads contained 6.2–7.2 HSV1 reads per million sequences analyzed (parts per million; ppm). The following two samples originated from patients with similar values of VZV DNA levels quantified by qPCR (1.9 and 3.9×10^5 Geq/ml). Despite equivalent levels a ten-fold difference in detected VZV reads was observed between sample two (15–16 ppm) and sample three (135–147 ppm). Sample 2 contained 272×10^6 white blood cells (WBC) per liter compared with sample 3 which contained 17×10^6 WBC per liter (Table 1). Thus, the difference in sensitivity was related to variations in the leukocyte composition in the samples.

To further test the sensitivity, two CSF samples containing JC polyomavirus (JCV), a DNA virus with a relatively small genome, were processed. One sample contained high virus levels (1.9×10^5 Geq/ml) and the other low virus levels (4.3×10^3 Geq/ml) (Sample 4–5). JCV DNA was readily detected in both samples ranging from 1757 to 2096 ppm in sample 4 and 40–57 ppm in sample 5.

In order to verify that the methodology was applicable for bacterial agents, we sequenced CSF from two patients with pneumococcal meningitis, diagnosed by cultivation and/or 16S rRNA gene Sanger sequencing (Sample 6–7). DNA from *Streptococcus pneumoniae* (*S. pneumoniae*) was classified with a range between 30,704 to 60,661 ppm (Sample 6), and 679–804 ppm (Sample 7). In addition to the bacterial samples, we included two CSF samples from patients with RNA viral enterovirus CNS infection (Sample 8–9). As expected, no DNA reads were identified. Enterovirus was, however, found using metagenomic RNA sequencing (Additional 1: Fig. 1).

Samples with co-infections, where EBV was detected along with a primary infectious agent (Enterovirus sample 9, VZV sample 10–11 and *Cryptococcus sp*. sample 12), were analyzed. EBV was not detected in sample

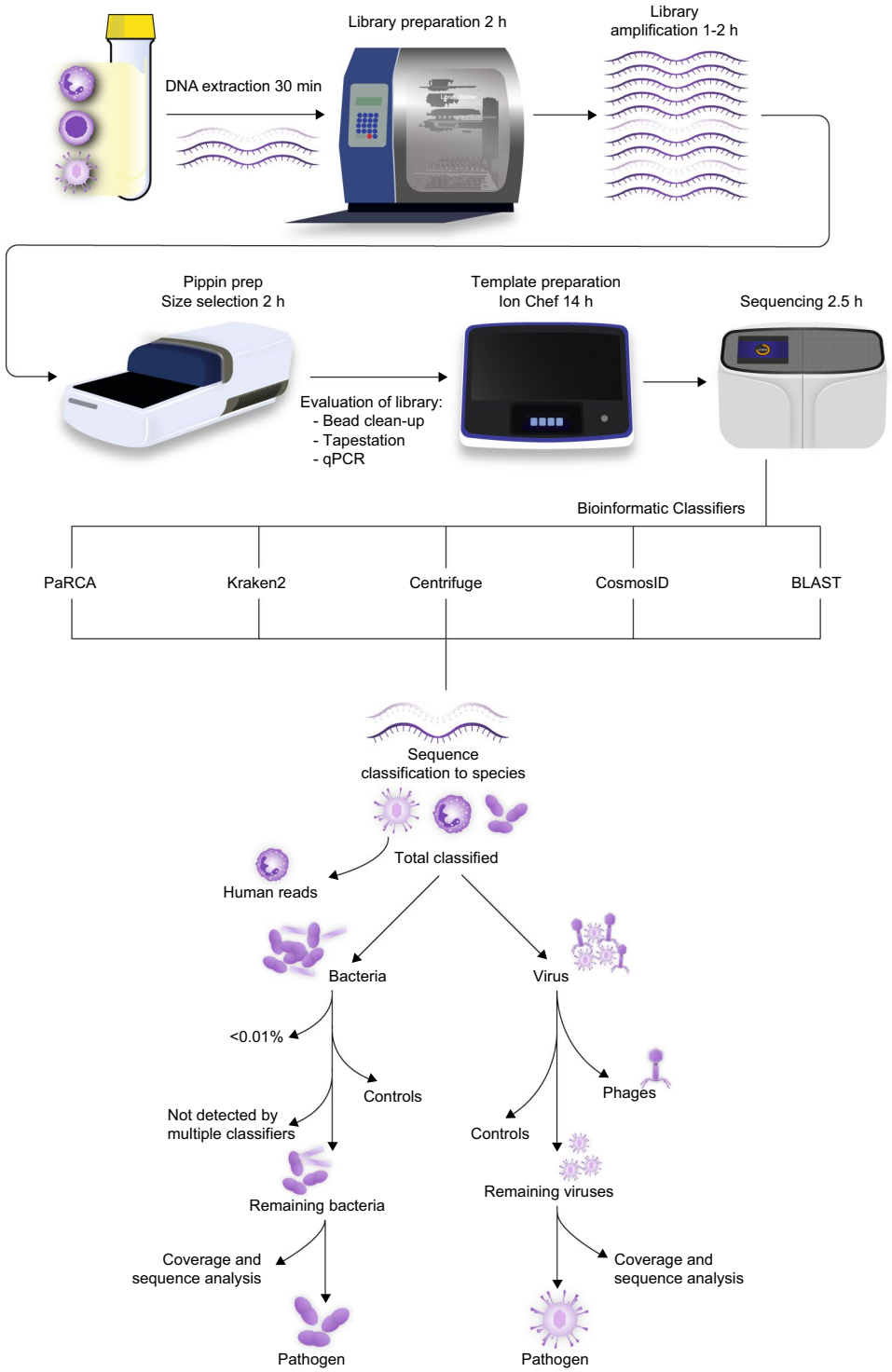

Figure 1. DNA metagenomic sequencing workflow. DNA from cerebrospinal fluid specimens, containing leukocytes and pathogens, was extracted and followed by library preparation and sequencing. Datasets generated by the Ion S5 were processed by four different bioinformatics classifiers to profile the microbiome. BLAST was used for verification. Flowchart for identification of pathogens by removing false positive species. Virus contaminants can be removed by comparison of sample datasets with controls by which environmental and bioinformatic misclassifications are identified. Phages can be disregarded as these viruses do not infect human cells. A final manual examination of remaining viral reads is required for coverage and sequence analysis. The bacterial contaminants were removed by applying a filter of cutoff value and comparison between classifiers and controls followed by a manual examination.

Sample	Verified pathogen	Clinical method	qPCR (Geq/ml)	PaRCA (reads)	Kraken2 (reads)	Centrifuge (reads)	CosmosID (reads)	BLAST (reads)	Calculated reads	Range (ppm)	Leukocytes ($\times 10^6$/l)
1	HSV1	qPCR	1.0×10^4	97	105	107	107	108	90	6.2–7.2	41
2	VZV	qPCR	3.9×10^5	213	219	223	211	213	365	14.9–16.0	272
3	VZV	qPCR	1.9×10^5	2196	2234	2251	2170	2197	3072	134.8–147.1	17
4	JCV	qPCR	1.9×10^5	23,766	24,018	24,190	22,318	23,847	N/A	1757–2096	N/A
5	JCV	qPCR	4.3×10^3	496	512	515	484	498	N/A	39.8–57.1	N/A
6	*S. pneumoniae*	Cultivation/16S rRNA	N/A	766,744	699,662	575,646	701,304	643,083	N/A	30,704–60,611	55
7	*S. pneumoniae* EBV	16S rRNA qPCR	N/A 3.7×10^2	12,988 –	11,762 –	12,511 –	12,277 –	12,274 –	N/A 0.1	679–804 Undet.	1064
8	Enterovirus	qPCR	6.6×10^4	–	–	–	–	–	N/A	Undet.	95
9	Enterovirus EBV	qPCR qPCR	5.8×10^4 4.1×10^2	– –	– –	– –	– –	– –	N/A 0.1	Undet. Undet.	814
10	EBV VZV	qPCR qPCR	1.9×10^3 4.7×10^3	10 7	9 7	9 7	8 7	9 7	2.5 4.5	0.8–1.1 0.7–0.8	181
11	EBV VZV	qPCR qPCR	5.0×10^1 2.9×10^3	– 15	– 15	– 15	– 12	– 15	0.1 5.5	Undet. 1.2–1.7	90
12	EBV Yeast sp.	qPCR Cultivation/filmarray	9.1×10^2 N/A	– –	– –	– –	– –	– –	0.2 N/A	Undet. Undet.	164
13	EBV	qPCR	1.9×10^3	81	85	82	79	82	20.5	6.7–7.5	26
14	EBV	qPCR	3.7×10^2	–	–	–	–	–	0.6	Undet.	253
15	EBV	qPCR	3.2×10^2	6	6	6	6	6	2.5	0.4–0.5	44
16	EBV	qPCR	2.7×10^2	232	228	225	213	223	18.5	21.2–22.8	4
17	EBV	qPCR	1.6×10^2	11	10	11	11	11	0.3	1.0–1.2	148
18	EBV	qPCR	1.6×10^2	–	–	–	–	–	N/A	Undet.	< 4
19	EBV	qPCR	8.1×10^1	–	–	–	–	–	0.6	Undet.	31
20	EBV	qPCR	5.0×10^1	–	1	1	–	1	0.99	0–0.1	14
21	EBV	qPCR	5.0×10^1	8	8	8	8	9	1.5	0.7–0.8	9

Table 1. Metagenomic sequencing pipeline results. Reads from each classifier from verified pathogen. Calculated reads in accordance with the presented algorithm N/A: leukocyte count missing for sample 4 and 5, leukocyte count for sample 18 is below reference value, calculation is not applicable for bacteria, fungi and RNA virus. *16S rRNA* 16S rRNA gene Sanger sequencing, *HSV1* Herpes simplex virus 1, *VZV* Varicella Zoster virus, *JCV* JC polyomavirus, *EBV* Epstein-Barr virus.

9 in accordance with the predicted sensitivity. The enterovirus RNA was as expected not detected in our DNA sequencing libraries. VZV and EBV were detected in sample 10, and only VZV was detected in sample 11. Neither yeast nor EBV DNA was detected in sample 12. The results were expected when the following equation was applied for calculating the sensitivity for each agent.

The theoretically expected number of pathogen reads was calculated using the Eq. (1) according to pathogen genome size (G_P), the diploid human genome size of 6.5 billion base pairs (G_H), pathogen copy according to PCR per milliliter (C_P), whole blood cell count per milliliter (C_H), and adjusted according to the volume (V), sequencing library size (L) and mappability in percent (M) to remove major contaminants.

$$Pathogen\,read = L/\left(\frac{C_H \times G_H \times V}{C_P \times G_P} \times M^{-1}\right)$$

Thus, in 0.3 ml CSF with a normal WBC count (5×10^3 per milliliter), containing one pathogen with a 1 million base pairs genome, a sequencing of library 10 million reads would produce one single pathogen read, provided that the mappability is > 95%.

We included additional nine CSF samples with low levels of EBV DNA (50–2000 Geq/ml) (Sample 13–21). With the exception of sample 13 (patient diagnosed with CNS Hodgkin's lymphoma type Post-Transplant Lymphoproliferative Disorder), and sample 16, where EBV was considered the cause of the symptoms, the EBV findings were clinically interpreted as benign incidental findings i.e. not the causative agent for the symptoms of infection. The EBV DNA detected in the majority of samples is likely to originate from latently infected B-lymphocytes recruited into the CSF. Despite the limitations for absolute quantification using qPCR and the stochasticity of distribution of low-level pathogen particles, with one exception the calculated reads correlated with the detected reads in the sequencing data (Table 1). In ten samples, more than 1 viral reads were expected and pathogen sequences were found in all samples (Fig. 2a). In eight samples where less than 1 read was expected to be found, EBV reads were only detected in one dataset (sample 17). Sixteen copies of EBV per milliliter was detected in sample 17 using qPCR and 11 reads were detected using metagenomic sequencing even though 0.3 reads were expected. The discrepancy between the calculation and sequencing results is most likely due to the stochastic distribution of the few viral particles in the sample. In sample 20, 0.99 reads were expected to be detected

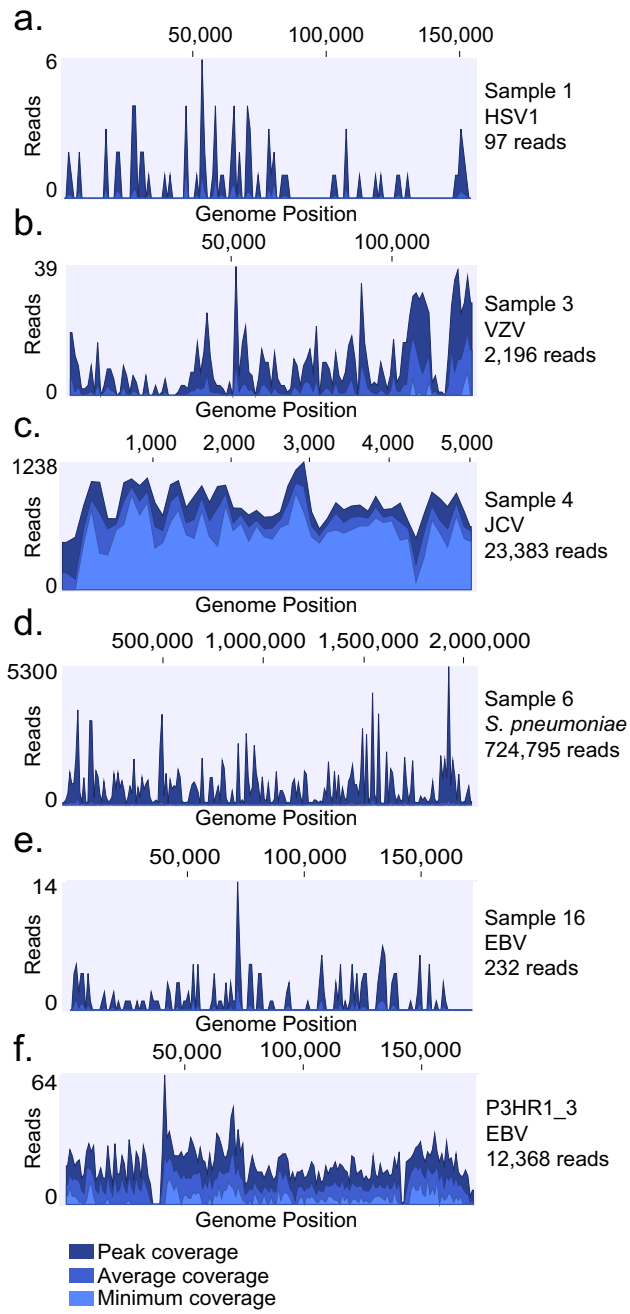

Figure 2. Pathogen genome alignment. Coverage density plot of sequencing reads from respective sample and control detected in PaRCA aligned to reference genomes of HSV1 (**a**), VZV (**b**), JCV (**c**), *S. pneumoniae* (**d**) and EBV (**e**, **f**). Number of reads (y-axis) at each nucleotide position of the genome (x-axis) depicted in blue. Dark blue represents peak, bright blue average and light blue minimum coverage for respective sections of the genome.

in the dataset and a single EBV-read was identified in two of the four classifiers (Kraken2 and Centrifuge). This read was further confirmed using BLAST. The WBC count in sample 18 was below the reference interval of the leukocyte cell counter and was therefore omitted.

All pathogen reads and control from PaRCA were mapped against the corresponding genome sequences using CLC genomics workbench (Fig. 2; Additional 1: Fig. 3). A dispersed distribution of the reads to the corresponding genomes was observed for all samples, except sample 10, where 5 of the 7 VZV reads (1 overlapping read) originate from a repetitive region within the genome and is therefore expected to be detected at a higher rate, and the last 2 reads map to a downstream gene (no overlap) (Additional 1: Fig. 3d). Each sequencing library was subjected to BLAST using the respective reference pathogen genome. The variation of the absolute number pathogen reads comparing the different classifiers detected was lower than 25% (Table 1).

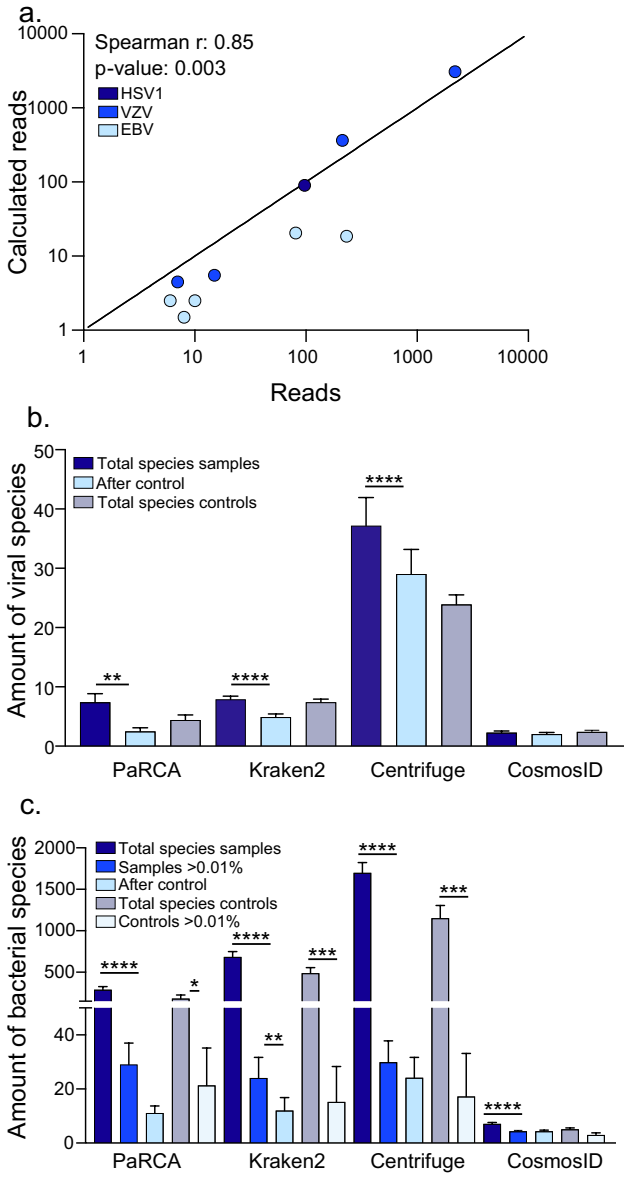

Figure 3. Calculated pathogen reads and detected pathogens in bioinformatic classifiers. Samples containing HSV1 (dark blue), VZV (blue) and EBV (light blue) with a calculated value of more than one read (**a**) were plotted against the number of reads detected by PaRCA. Regression line depicting a direct proportionality between the calculated and observed variables. The Spearman's rank correlation coefficient and p-value is indicated. Mean number ± SEM of viral (**b**) and bacterial species (**c**) classified in samples and controls using the different bioinformatic classifiers. Dark blue bars show the total number of species classified, bright blue bars show the amount of bacterial species over the fraction cutoff (≥ 0.01% of the dataset), light blue bars show number of species not removed using controls. Purple bars show controls and light gray show controls over the fraction cutoff (≥ 0.01%). Ordinary one-way ANOVA with Tukey's multiple comparisons, **p* value < 0.05, ***p* value < 0.01, ****p* value < 0.001, *****p* value < 0.0001.

Qualitative and quantitative detection of a known pathogen can thus reproducibly be carried out using the different types of bioinformatic classifiers. Furthermore, an estimation of sensitivity for pathogens can be generated for each sample which can guide the clinician whether the sequencing depth is sufficient to find a certain type of pathogen (Additional 4: Table 4). Notably however, each classifier produced diverse quantities of false positive hits.

False positive pathogens. The diversity of viral species detected in metagenomic sequencing libraries were relatively low and recurrent. PaRCA, Kraken2, Centrifuge and CosmosID identified 2–31, 5–13, 17–96 and 0–4 viral species in each sample respectively (Fig. 3b; Additional 1: Fig. 2a, Additional 4: Table 5). Many of the most abundant viral species identified were found in multiple samples (Fig. 3). Two samples (4 and 13)

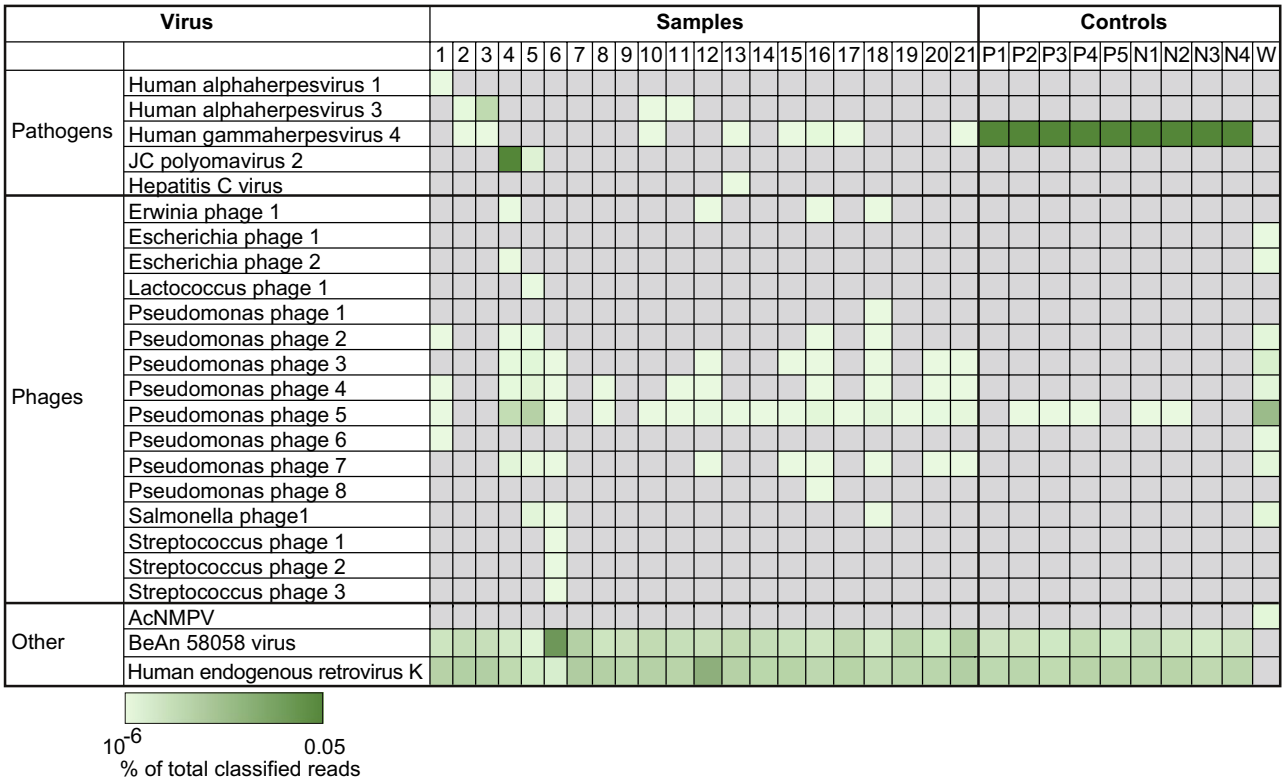

Figure 4. Viral species identified in datasets. Heatmap showing the ten most abundant viral species (y-axis) in each of the 21 samples and 10 controls (x-axis) detected using PaRCA. AcMNPV: Autographa californica multiple nucleopolyhedrovirus. Controls: P; P3HR1, N; Namalwa, W; water.

contained human viruses which were not detected in multiple samples and not a previously confirmed pathogen (see below).

The non-pathogen/EBV viral reads were either of human origin, misclassified or contaminations. Human endogenous retrovirus K was identified in all samples, except for the water control, which was expected as the reads originate from the human genome (Fig. 4 bottom; Additional 4: Table 5). Another ubiquitously detected virus was the BeAN 58,058 virus, which was detected in all samples, except for the water control. An additional BLAST examination identified these hits as human reads. Low levels of phage sequences known to infect bacteria from the *Enterobacteriales* order were detected in a few samples and in the water control, most likely derived from bacteria purified enzymes used in the various steps of library preparation. A conspicuous pseudomonas phage contaminant in sample 4, 5 and the water control are likely derived from a bacterial contaminant at one of the sequencing sites. Streptococcus phage species were detected in sample 6, from a patient with *S. pneumoniae* meningitis. Importantly, the most prominent viral species identified in patient samples were also present in the cell controls at similar levels and displayed a similar sequence identity and could therefore be discarded as a pathogen.

Compared with the relatively few viral agents detected by the classifiers, bacterial species were abundant; 61–712 bacterial species were identified using PaRCA, 370–1408 in Kraken2, 845–2826 in Centrifuge and 0–14 in CosmosID (Fig. 3c; Additional 1: Fig. 2b). Two samples originated from patients with a known *S. pneumoniae* meningitis (sample 6 and 7) and bacteria were detected at 69,088 ppm and 803 ppm respectively (PaRCA). With the omission of the positive samples 6 and 7, trace levels (3.4–18.2 ppm) of *S. pneumoniae* was ubiquitously detected in all samples. A known environmental contamination of *Pseudomonas* was detected in the majority of the samples. In two samples (4 and 5) *Pseudomonas* constituted 389,480 ppm (39%) and 590,195 ppm (59%) of the entire sequencing library respectively, while the prevalence in other samples were lower 6.6–75,279 ppm (0.0007–7.5%). A large fraction of the detected bacteria was still left when using previously suggested fixed cut-off at 100 ppm (0.01%) (Fig. 3c; Additional 1: Fig. 2b) and unlike the virus species the contaminants/misclassifications cannot be entirely removed using the control samples. However, when further applying an additional filter of comparison of the detected bacterial species between the three classifiers (PaRCA, Kraken2 and Centrifuge) only the known pathogen (*S. pneumoniae*) or environmental contaminants (*Pseudomonas* and *Escherichia coli*) was left. Similarly no eukaryotic species were found in all three classifiers.

Considering the ubiquitous presence of viral misclassifications and contaminants in samples as well as controls, a viral pathogen is easily identifiable, but requires additional analyses including read distribution and BLAST analysis, for verification in a clinical setting (discussed below). In contrast, the large number of bacterial species identified pose a bioinformatic challenge as the bacterial sequence can be derived from kit contaminants, lab environment or bioinformatic misclassifications which obscure the pathogen reads. As with the virus hits,

removal of bacterial contaminants using cell controls can efficiently remove the majority of species, but additional filters are required (Fig. 1).

Controls. Two types of controls, water and cell control, were tested for their ability to mirror the bioinformatic misclassifications and contaminations observed in samples. In the water control the dataset consisted of 99.6% bacterial sequences and 0.06% viral sequences (Additional 4: Table 5). The cell controls originating from EBV-transformed cancer cells had a composition more similar to the samples with 99.2–99.4% human sequences. The number of viral and bacterial strains detected in the water control was 12 and 568 respectively. In contrast the cell controls contain sequences ranging from 3–4 viral and 61–177 bacterial strains.

The viral strains in the water control were mainly of phage origin. In contrast the viral strains detected in the cell controls were similar to the CSF samples, mainly Human endogenous retrovirus K and BeAN 58,058 virus. Both cell lines originate from EBV-transformed cancer cells and harbors EBV DNA. The ppm-values of each cell line between sequencing runs were reproducible and no significant difference was found between the classifiers (Additional 1: Fig. 4, Additional 3: Table 6).

In the water control, 98% of the sequencing library consisted of reads from *Pseudomonas* and the second most abundant bacterial strain found was *Escherichia coli* (0.1%), which is to be expected as most enzymes are produced in this bacterial system. In contrast, none of the bacterial strains in the cell controls constituted more than 0.1% of the sequencing library.

Thus, the water control efficiently amplified the environmental and kit contaminants, but in contrast to the cell control did not find human misclassifications. Also, since the water control consisted entirely of contaminants, the absolute or proportionate content did not allow for a direct comparison with the patient samples. The cell control allowed for direct quantitative and qualitative subtraction of the majority of contaminants and putative pathogens were identified.

Unexpected virus findings. In sample 2 and 3 we identified 29–34 EBV reads in both samples in all classifiers (Additional 4: Table 5). The reads were dispersed throughout the genome and displayed minor sequence Pathogen detection in DNA metagenomic sequencing in CSF is mainly limited by the leukocyte count which affects the sensitivity and bioinformatically with the reference genome in accordance with previous EBV findings (Additional 1: Fig. 5a, b). Due to the limited sample volume we were unable to verify and quantify this finding using qPCR.

In sample 4 we identified three viruses which were unexpected, mastadenovirus, papillomavirus and torque teno virus (Additional 1: Fig. 5c–e). PaRCA identified 32 reads matching human mastadenovirus C (HAC), Kraken2 32 reads, Centrifuge 30 reads and CosmosID did not report any HAC sequences. The majority of reads, 25 out of 32 were 198 bp long, 5 reads were shorter and 2 were longer. BLAST-analysis showed that all reads shared the same 3'-end. Four reads had mismatches in comparison with the reference sequence. Considering the size and distribution of the reads our findings are most likely a laboratory amplicon contamination. Human papillomavirus (HPV) reads were detected in PaRCA (12 reads), Kraken2 (2 reads), but not by Centrifuge and CosmosID. Ten of the 12 reads were 105 bp long and the remaining two, 104 bp and 106 bp respectively. All reads aligned to the 3'-end of the virus genome in the L1 gene. Examination of BLAST results showed a high similarity with HPV98 with a one or two base pair mismatch. As above, considering the size and distribution of the reads our findings were most likely a laboratory amplicon contamination. CosmosID has an inbuilt function to filter out hits that are considered to be amplicons, therefore the software did not report these reads. Different strains of Anellovirus/Torque teno virus (TTV) were detected in the classifiers. PaRCA identified 75 reads, Kraken2 25 reads, Centrifuge 55 reads, while CosmosID did not detect any TTV reads. Five distinct consensus reads/contigs were formed from the 75 reads identified in PaRCA. Thirty-one reads formed a consensus read of 196 bp. BLAST analysis of this read displayed a 97% identity with TTV14, but only for 91 bp of the fragment. The remaining parts of the contig did not show any alignment with any viral species. The origin of this read is therefore unknown. BLAST analysis of the remaining 4 reads/contigs showed alignment (> 95% query cover and identity) to an Anellovirus isolate previously identified in metagenomics. The alignment showed an unusual coverage of the 5'-end of the genome and all the reads were aligned to the first half of the genome. The reason for this unusual coverage is unknown, but considering that TTV is widely detected in metagenomic sequencing and the multiple reads aligning to a clinical isolate it is probable that these four contigs/reads originate from the patient sample.

In sample 13, we detected 10 reads corresponding to hepatitis C virus (HCV) in PaRCA. Kraken2, Centrifuge and CosmosID detected 5, 6 and 6 reads respectively. The 10 reads were concentrated to the 5'-end of the genome, but spread within the initial half of the genome (Additional 1: Fig. 5f.). An analysis of the BLAST results showed alignment with HCV genotype 1. Synonymous mutations were found in multiple reads as well as gaps. Two reads had a fusion between sequences from different regions of the HCV genome. The sequence diversity indicates that the virus is from a patient, but the frameshift and fusion reads indicate that they are of an artificial origin. Also, the patient had undergone HCV serology analysis which was negative. Finally, considering that HCV is a RNA virus this finding is most likely a laboratory amplicon contamination.

Discussion

In this study we subjected 21 CSF samples from patients with suspected or confirmed CNS infection to metagenomic DNA sequencing. Pathogen detection accuracy and efficiency was evaluated using five bioinformatic tools. Using 14 samples with minute levels of EBV we concluded that the sensitivity of detection was mainly affected by leukocyte content in the samples and to lesser degree environmental contamination. Bioinformatic classifiers were essentially equally efficient in terms of sensitivity, but produced vastly different number of false positive hits, which inhibited efficient clinical pathogen identification. The removal of these false positive hits originating

from contaminants and bioinformatics classifications were alleviated by using an EBV-containing cell control which served as a positive as well as a negative control. A number of approaches have been suggested for how to identify a causative agent in clinical samples bioinformatically, e.g. by calculating the fraction of pathogen reads and/or an absolute number of reads[23,24]. However, applying this method, the majority of viral samples used in this study would be considered negative and/or contain a number of agents which would be considered falsely positive depending on the choice of classifier. The lower detection limit could be generalized and compared between studies/laboratories if the leukocyte count was provided. In a similar manner, a general quantification of viral content using ppm is an efficient, albeit insufficient reference point for comparison between studies[25,26]. An alternative would be an estimation of the number of pathogen copies per milliliter which can be calculated using our algorithm. Furthermore, it is evident that local contaminants greatly impact the sequencing library constitution. Therefore, it is necessary that findings in negative controls from each study be presented in its entirety. Major efforts have been invested in separating true pathogens from contaminations, e.g. using bioinformatic pipelines to find signatures of contaminants[27], which in combination with exogene controls and studentized residual approaches could determine whether a possible contaminant might be a causative agent[18]. Another approach is to use healthy patient control samples to build a dataset with expected ppm for a given finding, thus being able to calculate standard z-scores to remove contaminants[28]. Here, we calculated a ratio between samples and controls in a similar manner as previously described[13], which effectively removed the majority of background, leaving a manageable list of possible pathogens. Recently, guidelines from the European Society of Clinical Virology (ESCV) have been published with regards to both experimental procedures and bioinformatic pipelines, outlining the importance of careful consideration of each step in order to gain trustworthy results with minimal contamination[14,15,29]. We propose that information of the leukocyte content would also aid in the effort to determine whether a possible pathogen might be detectable.

Nine CSF-samples were identified at the clinic to only contain EBV, and we did not identify any additional pathogen, confirming the results from the clinic. Importantly, using our algorithm a lower detection limit could be determined for pathogens. Our bioinformatic classifier PaRCA, which uses a combination of single reads alignment and assemblies was able to detect more reads from HAC, HPV, TTV and HCV, but failed to detect the single EBV read in sample 20. Bioinformatic classifiers for clinical practice should not only quantify the pathogen reads, but also provide information of read distribution, sequence diversity and subtraction of environmental contaminants and bioinformatic misclassifications, facilitating pathogen detection as shown in this study. Novel pathogens will also require classifiers to detect diverse sequences, as well as enable investigation of sequences which might not classify completely to a genus. Our finding of a novel TTV strain shows that there is a large difference between bioinformatics classifiers' ability to identify divergent sequences.

In this study we have used material archived at − 20 °C, which impaired a proper RNA analysis due to degradation. However, with proper archiving at − 80 °C, it is manageable to perform RNA metagenomic sequencing[9]. Future studies using fresh CSF-samples where RNA integrity and quantity is measured may provide similar guidelines for RNA pathogen detection. We only included two verified bacterial CSF-samples in this study, one which was detected by culturing and 16S rRNA gene sequencing, and the second one detected by 16S rRNA gene sequencing. A limit of using metagenomic sequencing of CSF from bacterial meningitis patients is the high levels of leukocytes, but this may be compensated by the higher amount of bacterial nucleic acid compared with viral genomes, as previously observed[9]. Here, we applied a fraction cutoff for bacterial findings (> 0.01%) in order to decrease the amount of false positive bacterial species findings. This cutoff value should not be considered fixed and future studies with larger bacterial cohorts would provide additional guidelines for bacterial species identification.

We suggest that prior to clinical metagenomic DNA sequencing, an estimation of sequencing depth is made by adjusting it to the leukocyte content in the sample. Also, a pathogen-containing cell control sequenced at the same depth should be included in the same sequencing run in order to generate the same type of reproducible background. Bioinformatic processing should include a comparison between the pathogens detected in the cell control and the sample as well as between multiple classifiers. Further candidate pathogen reads should be confirmed by using BLAST and mapped against a reference genome to identify read distribution and sequence diversity. Even though true pathogen sequences should extremely rarely overlap, PCR duplicates can be present, especially in small sequencing libraries and/or samples that have undergone multiple cycles of amplification prior to sequencing. A comprehensive evaluation including a theoretical estimation on sensitivity of the metagenomics test as well as other clinical microbiological assays e.g. serology and PCR should assist the clinician in interpreting the final results.

Methods

All methods were carried out in accordance with relevant guidelines and regulations.

Sample collection. Included in this retrospective study were cerebrospinal fluid samples from patients with CNS symptoms of infection, in which the Department of Clinical Microbiology at Sahlgrenska University Hospital or the The Public Health Agency of Sweden previously had verified the infectious agent during 2015–2017. The sample cohort was mainly chosen to include samples with low levels of Epstein-Barr virus, but also a few samples with a variety of microorganisms (DNA/RNA virus, bacteria or fungi) determined by confirmatory testing using qPCR, cultures, 16S rRNA gene Sanger sequencing or FilmArray (Additional 2: Methods). The samples were stored at − 20 °C after clinical testing. The cell lines P3HR1 (HTB-62, American Type Culture Collection, ATCC, USA) and Namalwa (CRL-1432, American Type Culture Collection, ATCC, USA), were used as combined negative controls as well as positive controls, due to its inherent EBV genome. The controls were processed in parallel with the patient samples during all the laboratory steps.

Sample processing. For samples processed at the Department of Clinical Microbiology at Sahlgrenska University Hospital, total nucleic acid was extracted from 400 µl of cerebrospinal fluid using the MagNA Pure Compact Nucleic Acid Isolation Kit I (Roche Diagnostics, Indianapolis, IN, USA) on the MagNA Pure compact automated extractor. For samples processed at The Public Health Agency of Sweden, total nucleic acid was extracted from 200 µl of cerebrospinal fluid sample using the MagDEA® Dx SV (Precision System Science Co Ltd, Matsudo-city, Chiba, Japan) on the magLEAD® 12gC automated extractor (Precision System Science Co Ltd). DNA concentrations were determined using the Qubit Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) using the dsDNA HS Assay Kit (Thermo Fisher).

Library preparation and sequencing. DNA libraries were prepared according to the modified protocol for metagenomic samples, developed at the Public Health Agency of Sweden, using the Ion Xpress Plus Fragment Library Kit (Thermo Fisher) on the AB Library Builder System (Thermo Fisher). Samples were fragmented to 200 bp, followed by ligation of Ion P1 Adapter as well as Ion Xpress Barcode adapters. The protocol was adjusted to suit low-input samples (< 50 ng DNA) by using a reduced volume of P1 adapter and barcodes (0.5 µl). The libraries were amplified, selecting the number of amplification cycles according to the sample input concentration, varying between 14 to 20 cycles. Amplified libraries were size-selected choosing an optimal size range for each individual sample to ensure removal of small-sized PCR concatemers, varying between 100 to 320 bp (including adapters). Size selection was performed using the Pippin Prep platform (Sage Science, Beverly, MA, USA) with 2% Dye free Agarose Gel Cassette. Following visualization and an estimation of the concentration using the High Sensitivity D1000 DNA Kit on the Agilent 2200 TapeStation system (Agilent Technologies, CA, USA), the samples were pooled according to concentration. Subsequently, libraries were purified using Agencourt AMPure XP (Beckman Coulter, Brea, CA, USA). Finally, libraries were quantified using qPCR with the Ion Library TaqMan Quantitation Kit (Thermo Fisher) and the size estimated using High Sensitivity D1000 DNA Kit on Agilent 2200 TapeStation system (Agilent Technologies). For template preparation, libraries were pooled to a final concentration of 50 pM, if obtainable. For libraries with lower concentration than 50 pM, libraries were pooled to the available concentration. Thereafter, the Ion Chef Platform was used to ligate the libraries onto spheres using the Ion 540 Kit-Chef (Thermo Fisher). Following clonal amplification, libraries were loaded onto Ion 540 Chip and sequencing was performed on the S5 System (XL, Prime; Thermo Fisher) according to the manufacturer's protocol for 200 bp read length.

Ethical approval. The study design and methods including waiver of informed consent were approved by the Swedish Ethical Review Authority, Region Göteborg (191-18).

Bioinformatic analysis

Quality control. BAM-files were converted into fastq files using the Torrent Suite Software provided for the Ion S5 system. Reads were processed with FASTX toolkit[30] to fasta files. Fastqc was used to identify low-quality reads. Sequences were then subjected to the individual pipelines described below.

Pathogen detection for research and clinical applications (PaRCA). PaRCA is implemented as a fully automated one-click snakemake pipeline[31]. The pipeline can be configured to accept RNA-data, DNA-data or a combination of both from the same sample on a per-sample basis. Data from negative controls can also be added to provide a background for comparisons. PaRCA can use any protein and nucleotide database as input for classification, the ones used in this case are described below. The output of the pipeline is an aggregated HTML summary with Krona-graphs, sample statistics and pipeline performance. Classified reads for each identified species are also provided in the HTML interface using hyperlinks.

Databases were created using built-in tools in Kraken2 and Kaiju. Briefly, databases corresponding to bacteria, viruses and eukaryotes were created at DNA, RNA and protein level resulting in nine total k-mer databases. The viral databases consisted of all viral data in GenBank, the bacterial database consisted of the full Progenomes data[32] and eukaryotic databases were composed of the GenBank data for vertebrates, parasites and fungi. After download, the Progenomes database was continuously updated using scripts to reflect changes within the NCBI taxonomy. Reads were initially trimmed at both directions using BBDuk (BBMap 37.50) using an entropy mask of 0.9, trim quality of 16 and a minimum length of 40. Reads were corrected using Fiona (0.2.9) with id = 3 for substitution errors.

Reads were classified using Kraken2 and Kaiju by using individual databases[33]. Kraken2 results were filtered using the kraken-filter with a threshold of 0.15 for eukaryotes and 0.05 for viruses and bacteria (a higher threshold indicates higher stringency). Thresholds for Kaiju: score and minimum matches were set to 85.20 for eukaryotes, 80.18 for bacteria and 75.15 for viruses.

After initial classification and filtering, Kraken2 results were individually compared and reads with hits in multiple databases were evaluated based on k-mer score with the highest scoring match being retained for further downstream analysis. Kaiju scores were internally compared and the hit with the longest protein alignment was preserved. Reads with both Kraken2 and Kaiju hits were then compared and the lowest common ancestor of the two results was selected using mergeOutputs with "-c lowest" from the Kaiju package. Reads where the lowest common ancestor was a species designation were directly counted and saved while reads with a higher lowest common ancestor were further processed in the pipeline. Reads only classified by a single k-mer classifier were labelled as "singletons" and further processed.

Reads were ordered by taxonomic ID, which then were regressed through the taxonomic tree until either a genus-level or kingdom-level was reached. Reads without genus-level information or reads with a classification above genus level were stored separately for further analysis. After ordering into genus, all taxonomic IDs corresponding to a member of the genus were automatically downloaded from NCBI and corresponding accession identifiers were parsed from the NCBI accession dump file. Accession identifiers were then used to create a slice of the BLAST nt-database for that specific genus. Reads classified as belonging to the order "primates" were not processed further and received the taxonomic ID 9606 (Homo Sapiens).

Reads were analyzed in BLAST within the genus using a threshold of an e-value of 10^{-3} and the ten best hits were then retained. The ten results per read were parsed and the bit-score per taxon in the hits were aggregated. The taxon with the highest aggregate bit score was then selected as the putative taxon ID for the read. After taxon identification, results were merged and regressed in order to identify the species level classification of the putative taxon. If the kingdom level was reached before a species identification was found, the original taxon identifier was used in its place. Finally, any reads that were not successfully classified within a genus in the BLAST database creation step were collected and subjected to BLAST against the full NT-database with an e-value of $> 10^{-5}$ and a minimum query coverage of 20% as threshold, again the ten best hits were preserved. The results from both BLAST analyses were aggregated based on bit score and the resulting taxon ID regressed to species level if possible.

Classified reads were collected and presented using a krona-graph and tables in an html format. Tables were reorderable on name, taxonomic id and read count. Tables were also filterable, including wildcard functionality. FASTQ-files containing reads classified to an individual species and aggregates corresponding to kingdoms and unclassified reads were directly downloadable.

Kraken2. Kraken2-build download-library script was used to download the reference libraries for human, bacteria, virus, fungi, protozoa, plasmid and archaea. The human herpesvirus 4 (NC_009334) was added manually to the database for EBV mapping. Kraken2 was used to classify the reads using the default parameters with the dustmasker option.

Centrifuge. Centrifuge was used to classify the reads towards the default database including a manually added human herpesvirus 4 (NC_009334). Default parameters were used with the inbuilt quality control and repeat masker based dustmasking from NCBI tools[19]. In order to obtain reads from all pathogens included in this study, the total of both leaf and genus levels were incorporated from the Centrifuge reports, thus leading to higher amounts of total classified reads, however, since not all species were converted into the ETE3 toolkit, and some stops at genus level, this does not affect final results of classified pathogens.

CosmosID. Unassembled sequencing reads were directly analyzed using the commercially available genomic platform CosmosID to achieve identification of microbes at species level[34]. Each uploaded sample was searched and cleared from host sequences by the platform prior to analysis. CosmosID automatically filters out phages and amplicon-originated sequences.

BLAST. BLAST analysis was performed with reference genomes for the pathogens. The cutoff was set to $\geq 95\%$ sequence identity and an e-value of $\leq 10^{-3}$. Following standard steps for pre-processing reads, a BLAST search was performed with reads set as subjects and reference genomes set as queries. Reference genomes used were NC_001806 (HSV1), NC_001348 (VZV), NC_00196 (JCV), NC_003098 (*S. pneumoniae*), NC_007605 (EBV), NC_001405 (Human Mastadenovirus C; HAC), FM_955837.2 (Human Papillomavirus 98; HPV98), MH_649255.1 (Anellovirus), and NC_004102.1 (HCV).

Calculations and statistical analysis. Pathogen reads classified with PaRCA were uploaded to the CLC genomics workbench (Ver. 11, Qiagen) to perform and plot coverage analysis with the inbuilt tool "Map to reference". Read distribution over the reference genomes (listed under BLAST section) was evaluated with total number of pathogen reads in mind, and BLAST analysis was performed to confirm the read identity.

Classified sequences from Kraken2 and Centrifuge were visualized using Pavian[35]. Ratio between sample ppm and control ppm were calculated, where a ratio ≤ 10 was considered a contamination.

GrapPad Prism Ver. 7.0c was utilized to perform statistical analysis. Kruskal–Wallis test with Dunn's multiple comparison tests was applied to compare reproducibility through pipelines. Ordinary one-way ANOVA with Tukey's multiple comparisons was used to compare thresholds for background species removal. A p value ≤ 0.05 was considered significant.

Data availability

Available on the European Genome-phenome Archive. The PaRCA software is available on https://github.com/ClinicalGenomicsGBG/PARCA.

Received: 24 March 2021; Accepted: 4 February 2022
Published online: 01 March 2022

References

1. Granerod, J. *et al.* Causes of encephalitis and differences in their clinical presentations in England: A multicentre, population-based prospective study. *Lancet Infect. Dis.* **10**(12), 835–844 (2010).

2. Hong, H. L. *et al.* Clinical features, outcomes, and cerebrospinal fluid findings in adult patients with central nervous system (CNS) infections caused by varicella-zoster virus: Comparison with enterovirus CNS infections. *J. Med. Virol.* **86**(12), 2049–2054 (2014).
3. Okike, I. O. *et al.* Trends in bacterial, mycobacterial, and fungal meningitis in England and Wales 2004–11: An observational study. *Lancet Infect. Dis.* **14**(4), 301–307 (2014).
4. Maeda, E. *et al.* Spectrum of Epstein-Barr virus-related diseases: A pictorial review. *Jpn. J. Radiol.* **27**(1), 4–19 (2009).
5. Martelius, T. *et al.* Clinical characteristics of patients with Epstein Barr virus in cerebrospinal fluid. *BMC Infect. Dis.* **11**, 281 (2011).
6. Siddiqi, O. K. *et al.* Molecular diagnosis of central nervous system opportunistic infections in HIV-infected Zambian adults. *Clin. Infect. Dis.* **58**(12), 1771–1777 (2014).
7. Wilson, M. R. *et al.* Clinical metagenomic sequencing for diagnosis of meningitis and encephalitis. *N. Engl. J. Med.* **380**(24), 2327–2340 (2019).
8. Bowers, J. R. *et al.* Genomic analyses of acute flaccid myelitis cases among a cluster in Arizona provide further evidence of enterovirus D68 role. *MBio* **10**(1), e02262-18 (2019).
9. 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 (2019).
10. Gu, W., Miller, S. & Chiu, C. Y. Clinical metagenomic next-generation sequencing for pathogen detection. *Annu. Rev. Pathol.* **14**, 319–338 (2019).
11. Blauwkamp, T. A. *et al.* Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. *Nat. Microbiol.* **4**(4), 663–674 (2019).
12. Schlaberg, R. *et al.* Validation of metagenomic next-generation sequencing tests for universal pathogen detection. *Arch. Pathol. Lab. Med.* **141**(6), 776–786 (2017).
13. Miller, S. *et al.* Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome Res.* **29**(5), 831–842 (2019).
14. Lopez-Labrador, F. X. *et al.* Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure. *J. Clin. Virol.* **134**, 104691 (2021).
15. de Vries, J. J. C. *et al.* Recommendations for the introduction of metagenomic next-generation sequencing in clinical virology, part II: Bioinformatic analysis and reporting. *J. Clin. Virol.* **138**, 104812 (2021).
16. Bharucha, T. *et al.* STROBE-metagenomics: A STROBE extension statement to guide the reporting of metagenomics studies. *Lancet Infect. Dis.* **20**(10), e251–e260 (2020).
17. Strong, M. J. *et al.* Microbial contamination in next generation sequencing: Implications for sequence-based analysis of clinical samples. *PLoS Pathog.* **10**(11), e1004437 (2014).
18. Zinter, M. S. *et al.* Towards precision quantification of contamination in metagenomic sequencing experiments. *Microbiome* **7**(1), 62 (2019).
19. Kim, D. *et al.* Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**(12), 1721–1729 (2016).
20. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**(3), R46 (2014).
21. Andrusch, A. *et al.* PAIPline: Pathogen identification in metagenomic and clinical next generation sequencing samples. *Bioinformatics* **34**(17), i715–i721 (2018).
22. Kostic, A. D. *et al.* PathSeq: Software to identify or discover microbes by deep sequencing of human tissue. *Nat. Biotechnol.* **29**(5), 393–396 (2011).
23. Breitwieser, F. P., Lu, J. & Salzberg, S. L. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform.* **20**(4), 1125–1136 (2019).
24. Nooij, S. *et al.* Overview of virus metagenomic classification methods and their biological applications. *Front. Microbiol.* **9**, 749 (2018).
25. Tang, K. W. *et al.* The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat. Commun.* **4**, 2513 (2013).
26. Tang, K. W. & Larsson, E. Tumour virology in the era of high-throughput genomics. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**(1732), 20160265 (2017).
27. Davis, N. M. *et al.* Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**(1), 226 (2018).
28. Wilson, M. R. *et al.* Chronic meningitis investigated via metagenomic next-generation sequencing. *JAMA Neurol.* **75**(8), 947–955 (2018).
29. de Vries, J. J. C. *et al.* Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples. *J. Clin. Virol.* **141**, 104908 (2021).
30. FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/. Accessed 18 May 2020.
31. PaRCA. https://github.com/ClinicalGenomicsGBG/PARCA. Accessed 5 January 2022.
32. Progenomes2. http://progenomes.embl.de/. Accessed 18 May 2020.
33. Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **7**, 11257 (2016).
34. CosmosID. http://www.cosmosid.com/. Accessed 18 May 2020.
35. Breitwieser, F. P. & Salzberg, S. L. Pavian: Interactive analysis of metagenomics data for microbiome studies and pathogen identification. *Bioinformatics* **36**(4), 1303–1304 (2020).

Acknowledgements

DV, YT, SA, SHM and KWT were supported by Region Västra Götaland, Sweden. We thank the Bioinformatics Core Facility at the Sahlgrenska Academy for bioinformatics support.

Author contributions

This study was designed by H.E.J., M.L.K., S.B. and K.W.T. Cells were cultured by Y.T. Samples were selected by K.W.T., S.B. and D.V. Metagenomic sequencing was performed by M.L.K., S.B., and J.O. P.S., S.A. and S.H.M. executed bioinformatic analysis. Calculations were done by J.O., S.A. and K.W.T., D.V. and K.W.T. provided clinical expertise. Manuscript was written by J.O., S.B., H.E.J. and K.W.T. Figures and tables were prepared by J.O., Y.T. and K.W.T. All authors read and approved the final manuscript.

Funding

Open access funding provided by University of Gothenburg. This project was supported by funding from Region Västra Götaland, the Sahlgrenska University Hospital Fund C4A, FoU Laboratoriemedicin, the Konrad and Helfrid Johanssons Foundation, the Olle Engkvist foundation, the Längmanska foundation and the Wilhelm and Martina Lundgren foundation.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-07260-x.

Correspondence and requests for materials should be addressed to H.E.J.

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.