

OPEN ACCESS

EDITED BY

Lisa M. Sedger,
St.Vincent's Centre for Applied Medical
Research, Australia

REVIEWED BY

Yiqi Seow,
Technology and Research (A*STAR),
Singapore
Sofia Seabra,
New University of Lisbon, Portugal
Lora Pless,
University of Pittsburgh, United States

*CORRESPONDENCE

Josep Quer
✉ josep.quer@vhir.org

RECEIVED 13 January 2026

REVISED 19 March 2026

ACCEPTED 23 March 2026

PUBLISHED 02 April 2026

CITATION

Ibañez-Llagoña M, Colomer-Castell S,
Campos C, González-Camuesco Á,
Llauradó A, García-Larroy J,
Sánchez-Tejerina D, Rando-Segura A,
Andrés C, Esperalba J, Nadal P, Ferrer R,
Cortese MF, Tabernero D, Gregori J,
Riveiro-Barciela M, Ruiz-Cobo JC,
Ruiz A, del Barco E, Buti M, Goya M,
Antón A, Cano A, Juntas-Morales R and
Quer J (2026) Unveiling pathogens and
contaminants: refining metagenomics
for clinical diagnostics.
Front. Microbiol. 17:1786985.
doi: 10.3389/fmicb.2026.1786985

COPYRIGHT

© 2026 Ibañez-Llagoña,
Colomer-Castell, Campos,
González-Camuesco, Llauradó,
García-Larroy, Sánchez-Tejerina,
Rando-Segura, Andrés, Esperalba, Nadal,
Ferrer, Cortese, Tabernero, Gregori,
Riveiro-Barciela, Ruiz-Cobo, Ruiz, del
Barco, Buti, Goya, Antón, Cano,
Juntas-Morales and Quer. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which does
not comply with these terms.

Unveiling pathogens and contaminants: refining metagenomics for clinical diagnostics

Marta Ibañez-Llagoña^{1,2,3}, Sergi Colomer-Castell^{1,3,4},
Carolina Campos^{1,3,4}, Álvaro González-Camuesco^{1,2},
Arnau Llauradó^{2,5,6,7}, Jorgina Garcia-Larroy¹,
Daniel Sánchez-Tejerina^{2,5,6,7}, Ariadna Rando-Segura^{2,8},
Cristina Andrés⁸, Juliana Esperalba^{8,9}, Patricia Nadal⁸,
Roser Ferrer^{4,10}, Maria Francesca Cortese^{1,3,8},
David Tabernero^{1,3,8}, Josep Gregori¹, Mar Riveiro-Barciela^{1,3},
Juan Carlos Ruiz-Cobo^{1,2,3}, Agustin Ruiz^{11,12,13}, Ester del
Barco^{14,15}, Maria Buti^{1,2,3}, Maria Goya^{14,15}, Andrés Antón^{8,9},
Amanda Cano^{11,12}, Raúl Juntas-Morales^{2,5,6,7} and
Josep Quer^{1,2,3,4*}

¹Liver Diseases-Viral Hepatitis, Liver Unit, Vall d'Hebron Institut de Recerca (VHIR), Instituto de
Investigación Sanitaria Hospital Universitari Vall d'Hebron (IIS IR-HUVH), Barcelona, Spain, ²Medicine
Department, Universitat Autònoma de Barcelona (UAB), Bellaterra, Spain, ³Centro de Investigación
Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Madrid, Spain, ⁴Biochemistry
and Molecular Biology Department, UAB, Bellaterra, Spain, ⁵Neuromuscular Diseases Unit, Neurology
Department, IIS IR-HUVH, Vall d'Hebron Barcelona Hospital Campus, Barcelona, Spain, ⁶Peripheral
Nervous System Group, VHIR - IIS IR-HUVH, Vall d'Hebron Barcelona Hospital Campus, Barcelona,
Spain, ⁷European Reference Network on Rare Neuromuscular Diseases (ERN EURO-NMD), Barcelona,
Spain, ⁸Microbiology Department, VHIR - IIS IR-HUVH, Vall d'Hebron Barcelona Hospital Campus,
Barcelona, Spain, ⁹Centro de Investigación Biomédica en Red de Enfermedades Infecciosas
(CIBERINFEC), Madrid, Spain, ¹⁰Biochemistry Department, Vall d'Hebron Hospital Universitari, Vall
d'Hebron Barcelona Hospital Campus, Barcelona, Spain, ¹¹Research Center and Memory Clinic, ACE
Alzheimer Center Barcelona, Universitat Internacional de Catalunya, Barcelona, Spain, ¹²Ciber
Neurodegenerative Diseases (CIBERNED), Madrid, Spain, ¹³Department of Microbiology, Immunology
and Molecular Genetics, Long School of Medicine, Glenn Biggs Institute for Alzheimer's &
Neurodegenerative Diseases, The University of Texas Health Science Center, San Antonio, TX, United
States, ¹⁴Maternal-Fetal Medicine Unit, Department of Obstetrics, VHIR - IIS IR-HUVH, Vall d'Hebron
Barcelona Hospital Campus, Barcelona, Spain, ¹⁵Department of Pediatrics, Obstetrics and
Gynecology, and Preventive Medicine and Public Health, Universitat Autònoma de Barcelona
(Bellaterra), Bellaterra, Spain

Introduction: Shotgun metagenomic sequencing (mNGS), an untargeted
approach that sequences all nucleic acids in a sample, has emerged as a powerful
tool for pathogen detection and genome characterization. However, its imple-
mentation in clinical diagnostics remains limited due to technical challenges such
as contamination and reduces sensitivity, especially in low-biomass samples.

Methods: We applied mNGS to 144 clinical samples representing chronic infec-
tions, acute infections, and respiratory co-infections. To address contamination,
we established a framework integrating negative controls, lab-specific contami-
nant watchlists, and computational filtering. Viral detection performance and
genome recovery were assessed across sample types and viral loads.

Results: Viral load was shown to be the primary determinant of sensitivity,
with reliable recovery achieved only at higher titers. Our framework substan-
tially improved contamination management, reducing false-positive signals and

enhancing viral genome recovery. mNGS enabled the detection of clinically relevant co-infections and refined viral classification beyond targeted diagnostics, while also revealing the substantial risk of spurious detections in the absence of contamination-aware workflows.

Discussion: These findings define practical sensitivity thresholds for clinical mNGS and underscore the need for contamination-aware workflows, particularly for low-biomass samples, while providing an open-source contaminants watchlist that enhances reliability and utility of clinical metagenomics.

KEYWORDS

coinfection, diagnostics, metagenomics, sequencing, virology

1 Introduction

Shotgun metagenomics (mNGS) is a next-generation sequencing (NGS) approach that enables comprehensive analysis of all nucleic acids (both DNA and RNA), within a given sample (Quer et al., 2022; Ibañez-Lligoña et al., 2023). By using a non-targeted strategy, mNGS allows for unbiased sequencing of microbial and host genetic material, facilitating the detection and characterization of a wide range of microorganisms, including viruses, bacteria, fungi, archaea, and parasites (Chrastek et al., 2022; Parras-Moltó et al., 2018; Iyer and Damania, 2020; Xie et al., 2023; Zhao et al., 2024; Miao et al., 2018; Su et al., 2024). This hypothesis-free approach is particularly well suited for identifying novel or unexpected pathogens and understanding microbial diversity and dynamics in complex clinical settings (Ibañez-Lligoña et al., 2023). Metagenomics has proven to be valuable in a wide range of applications including study of microbial community composition, pathogen evolution, antimicrobial resistance, and the discovery of novel microorganisms through whole-genome sequencing (Quer et al., 2022; Chiu and Miller, 2019; von Fricken et al., 2023).

Over the last decade, mNGS has been introduced into the clinical sphere (Comas et al., 2020), where it offers several advantages over conventional diagnostics (Chiu and Miller, 2019; Gu et al., 2019). Traditional methods such as culture, PCR or serological tests rely on prior knowledge of the target organism and fail to identify unculturable, unknown, or unexpected pathogens (Ibañez-Lligoña et al., 2023; Schloss and Handelsman, 2005). Moreover, whole-genome sequencing enables functional characterization, identification of resistance mutations and virulence factors, thereby supporting both diagnosis and informed therapeutic decision-making.

Accurate disease diagnosis is essential for effective treatment, as misdiagnoses causes delays, increased mortality and higher healthcare costs. Acute infections remain particularly challenging, with 50–60% of hospitalized patients being discharged without an identified cause. In emerging infections and chronic conditions, low-abundance or resolved infections often go undetected by conventional methods. Since syndromic presentations can result from multiple pathogens, a high-throughput approach that simultaneously identifies all potential agents in one assay is needed (Edridge et al., 2019). Moreover, mNGS could be used for prompt detection of emerging, re-emerging, or novel pathogens is essential for timely outbreak containment (Quer et al., 2022).

Despite its potential, clinical implementation of mNGS remains limited due to technical, analytical and interpretative challenges (Ibañez-Lligoña et al., 2023). Contamination is one of the major technical barriers, which is exacerbated by the technique's high sensitivity (Jurasz et al., 2021). Exogenous nucleic acids can be introduced at

multiple stages of the workflow, from sample collection to sequencing (Fierer et al., 2025). Common sources include reagents and consumables (collectively known as the "kitome" [Olomu et al., 2020]), laboratory environments, and operator handling (Lou et al., 2023). These contaminants can lead to false-positive results or obscure genuine microbial signals. This is particularly critical in low-biomass samples, where the actual amount of microbial genetic material is very small compared to the host or background content, therefore enabling contaminants and host-derived reads dominate the sequencing output (Jurasz et al., 2021; Fierer et al., 2025; Minich et al., 2019).

Several strategies have been developed to mitigate the risk of contamination, including physical decontamination methods (e.g., UV treatment, DNase digestion), bioinformatic filtering, and the use of negative controls (Fierer et al., 2025; Corless et al., 2000). However, no single approach has proven to fully remove background noise. In addition, few studies have systematically catalogued the composition and behaviour of contaminants across large clinical datasets or assessed their impact on diagnostic performance. This knowledge gap complicates efforts to set detection thresholds or confidently interpret results from low-biomass samples. In this study, we aim to address key limitations in clinical mNGS workflows by characterizing recurrent contamination and evaluating the natural viral sensitivity across 144 clinical samples and 18 methodological controls.

2 Materials and methods

2.1 Sample collection

A total of 144 clinical samples were collected from patients across various diagnostic groups representing diverse diagnostic categories and subjected to shotgun metagenomic sequencing (Table 1). All samples were leftover material obtained from routine clinical diagnostic testing. Samples were completely anonymized. Human data of the participants related to sex, gender, race, ethnicity, social grouping, and other social variables are not relevant to the scope of the analysis. Our research has been approved by the Vall d'Hebron Barcelona Hospital Campus Ethics Committee under the following approval numbers: PR(AG)429–2021 and PR(AG)287–2022 for neuropathy samples, PR(AG)259–2020 for the respiratory virus samples, PR(AG)118–2021 for HEV samples, and PR(AMI)437–2023 for pregnant women samples. Written informed consent was obtained from all patients for the collection and research use of their biological samples.

Among these, samples from patients with Guillain-Barré Syndrome (GBS) were also analyzed, which included 26 cerebrospinal fluid (CSF)

TABLE 1 Distribution of the 144 samples analyzed, categorized by sample type and indicating paired samples.

Condition	Sample type	Article code	Number of samples	Paired samples
Alzheimer's disease	CSF	ALZ-CSF	39	NA
Guillain-Barré Syndrome	CSF	GBS-CSF	26	21
	Serum	GBS-S	21	
Neuralgic amyotrophy	CSF	NA-CSF	1	1
	Serum	NA-S	1	
Preterm birth	Plasma	PB-P	4	4
	Amniotic fluid	PB-AF	6	
Positive controls	Nasopharyngeal swab	PC-NFS	11	NA
	Nasopharyngeal exudate	PC-NFE	4	
	Plasma	PC-P	10	
Acute HEV	Amniotic fluid	HEVA-AF	1	2
	Feces	HEVA-F	1	
	Plasma	HEVA-P	2	
Chronic HEV	Plasma	HEVC-P	17	NA

The codification depending on the condition and type is added for easy interpretation of the results. AF, amniotic fluid; ALZ, Alzheimer's disease; CSF, cerebrospinal fluid; F, feces; GBS, Guillain-Barré Syndrome; HEV, Hepatitis E virus; HEVA, Acute Hepatitis E virus infection; HEVC, Chronic Hepatitis E virus infection; P, plasma; PB, preterm birth; PC, positive control; NA, neuralgic amyotrophy; NFE, Nasopharyngeal exudate; NFS, nasopharyngeal swab.

and 21 serum specimens, of which 21 of them form matched CSF-serum pairs. Both serum and CSF samples were collected in the context of acute symptomatology for suspected GBS cases, with posterior diagnostic confirmation of included cases. Additionally, 1 matched CSF-serum pair was obtained from a case of neuralgic amyotrophy, another acute inflammatory peripheral nervous system disorder. CSF was obtained through a lumbar puncture performed for clinical diagnostic purposes, with a surplus of 2 to 4 mL retained for research. The CSF samples were then frozen at -80°C for preservation. For serum collection, whole blood was drawn into 9 mL blood tubes and centrifuged at 3500 rpm at 4°C for 15 min. The resulting serum was aliquoted into 1.5 mL tubes. Serum was then frozen at -80°C for preservation.

Additionally, 39 CSF samples were collected from individuals diagnosed with Alzheimer's disease (AD), following the methodology described in the [Cano et al. \(2024\)](#). Six amniotic fluid samples and four plasma samples, including four matched pairs, were collected from pregnant women in the clinical context of suspected intra-amniotic infection (chorioamnionitis). Gestational ages at sampling ranged from 16 to 33.6 weeks. Upon hospital admission, an amniocentesis was performed under ultrasound guidance for diagnostic purposes. Using continuous ultrasound monitoring, a sterile needle was inserted via transabdominal puncture to aspirate a sample of amniotic fluid (AF) under strict aseptic conditions.

In addition to the mentioned clinical samples, a set of positive controls was included consisting of clinical specimens with confirmed viral infections using conventional clinical microbiology methods. Eleven nasopharyngeal swabs, four nasopharyngeal exudates and ten plasma samples were obtained from cases of suspected acute infection in patients admitted to the hospital or emergency department. Moreover, one stool, one amniotic fluid and two plasma samples were collected from a patient which had tested positive for a HEV acute infection. Furthermore, 17 plasma samples were obtained from a patient with a chronic HEV infection.

Samples were classified as low-biomass or high-biomass according to their anatomical origin. Low-biomass samples were defined as

specimens derived from physiologically sterile body sites, including cerebrospinal fluid, plasma, serum, and amniotic fluid, which are not expected to harbour a resident microbiome under basal conditions. High-biomass samples comprised specimen types known to contain abundant microbial communities, such as stool and nasopharyngeal swabs and exudates ([Fierer et al., 2025](#)). Blank negative controls consisted of DNase and RNase-free sterile water (ThermoFisher Scientific, Waltham, MA, USA), which is routinely used for nucleic acid extractions and RT-PCR-Nested reactions.

2.2 Nucleic acid extraction, library preparation and Illumina sequencing

For nucleic acid extraction from nasopharyngeal swab and exudates, the STARMag 96 $\times$ 4 Universal Cartridge Kit was used on the Microlab STARled automated platform (Seegene, South Korea). In contrast, RNA and DNA from amniotic fluid, cerebrospinal fluid, plasma, serum and stool samples were extracted using the QIAamp MinElute Virus Spin extraction kit (QIAGEN, Hilden, Germany), omitting the addition of the RNA carrier. Next, sequencing libraries were prepared through the TruSeq Stranded total RNA library kit (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. Unique dual indices (IDT for Illumina – TruSeq RNA UD Indexes v2) were ligated to each sample to enable multiplexing. Following library quantification and normalization to 4 nM, libraries were pooled prior to sequencing.

Additionally, two types of negative methodological controls were included in the workflow: extraction controls (CE), added during nucleic acid extraction (sterile water), and library controls (CL), added during library preparation and processed in parallel with the clinical samples throughout the workflow. A total of eighteen “blank” controls were included across different sequencing runs to assess and control for inter and intra-run sequencing variations.

Library quality was assessed using the KAPA Library Quantification Kit (Roche Applied Science, Pleasanton, CA, USA) by

RT-qPCR in the Light Cycler 480 instrument in combination with the TapeStation 4,200 system (Agilent, Santa Clara, CA, USA) using the D1000 ScreenTape Assay. Finally, libraries were sequenced on either the NextSeq2000 or NovaSeq6000 platforms (Illumina, San Diego, CA, USA), with PhiX v3 (Illumina, San Diego, CA, USA) used as internal control for sequencing.

Sequencing generated a total of 9.81×10^{10} reads across all eleven runs. Two of these runs were performed in a NovaSeq6000 platform, while the other nine runs were done in the NextSeq2000 platform. Libraries sequenced on the NextSeq2000 platform produced a median of 9.53×10^7 reads per sample (3.4×10^6 – 1.54×10^8), whereas NovaSeq6000 runs yielded a median of 1.61×10^8 reads per sample (6.67×10^7 – 4.51×10^8 reads).

2.3 Cross-contamination control experiment

Additionally, three experiments were carried out to assess the putative cross-contamination that occurred during the process of library preparation and/or sequencing. For that, a NextSeq2000 (Illumina, San Diego, CA, USA) was loaded with a pool containing eight water “blank” controls, a previously produced in-lab HCV clone and one RNA coming from HEVC-P-2 sample which underwent library preparation through TruSeq Stranded total RNA library kit (Illumina, San Diego, CA, USA).

2.4 Bioinformatic analysis

Sequencing data was retrieved from the sequencing platforms in either FASTQ or BCL format for all clinical samples and corresponding negative controls. When required, BCL files were converted to FASTQ format using the bcl2fastq software (Illumina, San Diego, CA, USA). Then, samples underwent an initial quality control (QC) assessment.

QC analysis began with the generation of quality reports using FastQC (Babraham Bioinformatics, 2010) and MultiQC (Ewels et al., 2016). Next, raw reads were trimmed using Trimmomatic (Bolger et al., 2014) to remove adapter sequences, low-quality bases, and short reads. Duplicate sequences were removed using the clumpify tool from BBMap suite (Bushnell, 2014). To minimize the influence of technical contamination, BBDuk from BBMap suite (Bushnell, 2014) was used to filter out reads present in matched extraction (CE) and library (CL) controls.

Human reads were identified by mapping against the human reference genome (GRCh38) using Bowtie2 (Langmead and Salzberg, 2012) and subsequently removed from the dataset. The remaining non-human reads were subjected to taxonomic classification using Kraken2 (Wood et al., 2019) with the PlusPF database (RefSeq database for archaea, bacteria, viruses, plasmids, human, UniVec_Core, protozoa and fungi; accessed in April 2025). Taxonomic results were then filtered using an in-house script to remove known environmental and reagent-derived contaminants, based on both internal negative controls and previously published contaminant lists. *De novo* assembly of high-quality, non-human reads was performed using MEGAHIT (Li et al., 2015). Assembly quality was evaluated by remapping reads back to the assembled contigs. Assembled contigs were further classified taxonomically using Kraken2 (Wood et al., 2019). To validate the presence of specific pathogens identified through taxonomic classification and to assess genome coverage, reference-based mapping was

conducted on selected positive samples using Bowtie2 (Langmead and Salzberg, 2012).

All downstream analyses were conducted in R (R Core Team, 2022) (v 4.2.5). Data processing and visualization, including Principal Coordinate Analysis, bar plots, line graphs, and statistical summaries, were performed using the tidyverse (Wickham et al., 2019), ggplot2 (Wickham, 2016), and other relevant R packages. Custom scripts were used to generate read-retention plots across pipeline steps, exploring taxonomic profiles and compare microbial compositions between sample groups and control types. In parallel, decontam (Davis et al., 2018) was used to identify contaminants in our data.

To contextualize the presence of recurrent contaminant genera in clinical samples, the proportion of non-human reads assigned to genera included in the defined contaminant watchlist was calculated for each sample using raw reads counts prior to normalization. Moreover, rarefaction (saturation) analyses were performed in host-depleted, pre-filtered genus-level files to assess sequence depth adequacy across sample types with the use of the R package vegan (Oksanen et al., 2025).

3 Results

3.1 Characterization of recurrent contaminants in negative controls

Negative controls showed highly consistent contaminant profiles dominated by recurrent taxa, with no significant differences between extraction vs. library controls. Notably, the next step involving host read depletion altered overall compositional clustering patterns and reduced the total read counts, indicating that high host content dominates Bray–Curtis dissimilarity when present and can obscure microbial-specific variation (Supplementary Figures S1, S2).

To assess variability in community composition across control types, we performed PERMANOVA and beta-dispersion analyses after each processing step. No significant differences were detected between extraction and library controls at any stage (PERMANOVA p -value > 0.05, dispersion p -value > 0.05). Although statistical testing did not reveal any significant difference in beta dispersion at the human-read-removed stage, there is an apparent difference in spread between extraction and library controls observed in the dispersion boxplot (Supplementary Figure S3).

After filtering, a contaminant watchlist was generated by identifying taxonomic families (Supplementary Table S1) and genera (1271) (Supplementary Table S2) that were prevalent in at least half of all negative controls and had a counts per million (CPM) > 0.5 (Table 2). From this list, 81 genera such as *Propionibacterium*, *Gammaretrovirus* or *Flavobacterium*, matched previously published contaminants (Jurasz et al., 2021; Lou et al., 2023; Asplund et al., 2019; Piro and Renard, 2022; Zinter et al., 2019; Salter et al., 2014; Hornung et al., 2019; Miller et al., 2019) (Supplementary Table S3). Moreover, we extracted specific genera according to methodological control type (Supplementary Tables S4, S5).

Application of this contaminant watchlist reduced the total number of genera to be assessed for downstream analysis in clinical samples by 44.7%, substantially simplifying clinical reporting. After host removal, the proportion of non-host reads classified as recurrent contaminants from the watchlist varied across sample types. Among

TABLE 2 Top 15 most prevalent and abundant genera detected across all methodological controls.

Genus	NCBI taxonomy ID	Prevalence (%)	Mean CPM	Median CPM	Interquartile range of CPM
<i>Phyllobacterium</i>	28100	100.00	434929.23	746526.62	770279.24
<i>Burkholderia</i>	32008	100.00	166363.15	7837.61	360049.24
<i>Escherichia</i>	561	100.00	40490.29	4817.07	34218.42
<i>Alcaligenes</i>	507	100.00	18907.37	774.93	39799.55
<i>Pseudomonas</i>	286	100.00	16802.39	3018.45	4813.42
<i>Micrococcus</i>	1269	100.00	13736.17	1327.76	6938.39
<i>Cutibacterium</i>	1912216	100.00	11400.18	3053.41	22247.96
<i>Sphingomonas</i>	13687	100.00	10633.93	1176.37	3583.25
<i>Moraxella</i>	475	100.00	10615.59	671.47	9709.87
<i>Homo</i>	9605	100.00	9350.71	4009.78	4866.26
<i>Mesorhizobium</i>	68287	100.00	9274.51	5730.84	15776.96
<i>Corynebacterium</i>	1716	100.00	7134.13	1714.06	8339.69
<i>Shewanella</i>	22	100.00	6549.39	424.87	10470.74
<i>Paraburkholderia</i>	1822464	100.00	5699.98	670.65	3090.06

The table includes genus name, NCBI Taxonomy ID, and abundance statistics: mean counts per million (CPM), calculated from raw genus-level read counts and normalized to the total number of reads per sample. Summary statistics (mean, median, and interquartile range (IQR)) are shown across all negative controls.

low-biomass samples, amniotic fluid showed the highest mean percentage of contaminant reads (98.6%), followed by serum (97.7%), cerebrospinal fluid (96.9%), and plasma (94.4%).

High-biomass samples showed greater variability. Nasopharyngeal exudates (97.3%), nasopharyngeal swabs (92.6%) displayed contaminant proportions comparable to low-biomass samples, whereas the fecal sample showed a substantially lower proportion of contaminant reads (48.9%) consistent with a higher endogenous microbial content (Supplementary Tables S6, S7).

To explore sample-control relationships, PCoA was performed on 151 samples (including replicates) and 18 blank negative controls after each pipeline step. Samples were stratified by biomass level (low-biomass samples, high-biomass samples). Moreover, rarefaction (saturation) analyses were performed stratified by clinical sample type to evaluate sequencing depth adequacy (Supplementary Figures S4–S6). Across sample types and methodological controls, curves demonstrated early plateauing behaviour indicating that sequencing depth was sufficient to capture the majority of the detectable taxa within each matrix.

In early processing steps (after the removal of low-quality sequences), low-biomass samples clustered tightly with controls, and PERMANOVA revealed no significant compositional differences (PERMANOVA p -value > 0.05, dispersion p -value > 0.05, Figure 1a), indicating the domination of contamination, due to limited true microbial signal. For decontamination, we employed BBDuk (Bushnell, 2014)-based read subtraction using negative controls as reference. This led to partial separation of low-biomass samples from controls in the Bray-Curtis space (Figure 1b), although the PCoA still showed overlapping. PERMANOVA confirmed a statistically significant difference between low-biomass samples and methodological negative controls ($R^2 = 0.053$, p -value < 0.001). However, beta dispersion analysis indicated a significantly higher within-group variability ($F = 25.0$, p -value < 0.001), suggesting that heterogeneity in dispersion may partially contribute to the observed group differences. Conversely, PCoA of high-biomass samples and controls showed a clear

separation from controls as early as the trimming step (PERMANOVA p -value < 0.05) (Figure 1c). This trend continued across the subsequent pipeline stages, and by the last step, separation was pronounced, with PCoA1 explaining 36.4% of the variance (Figure 1d). In contrast, low-biomass samples clustered tightly with controls across steps.

To further reduce background noise, we applied prevalence-based filtering using decontam (Davis et al., 2018), alongside our contaminant watchlist for interpretative purposes to flag ambiguous or low abundance taxa without risking the exclusion of true biological signals (Supplementary Figure S7). Despite elevated group heterogeneity (beta dispersion p -value < 0.05), the total retained microbial abundance per sample was significantly higher in clinical samples compared to negative control (Wilcoxon p -value < 0.001), supporting the presence of true biological content distinct from background contamination.

3.2 Assessment of cross-contamination and index hopping

To evaluate cross-contamination, nine negative controls (sterile water), one of which was spiked with a hepatitis C Virus (HCV) clone were sequenced. A hepatitis E virus (HEV)-positive plasma sample was included in the same batch in three separate sequencing runs. Despite stringent protocols, HCV reads were detected in other samples in all three sequencing runs at significantly lower read counts (between 2 to 161 reads) compared to the HCV sample. HEV reads appeared only in the HEV samples in two out of the three replicate sequencing runs. However, HEV reads appeared in the HCV-spiked sample at very low read counts in one of the batches (Figure 2a).

To further investigate these low-level detections, we analyzed shared tile and index combinations across samples to identify potential index misassignment events in the three sequencing runs. This analysis revealed pairwise misassignment events across samples in the three sequencing runs, with a pattern of dual-index sharing consistent with index hopping. Most sample pairs showed minimal read transfer

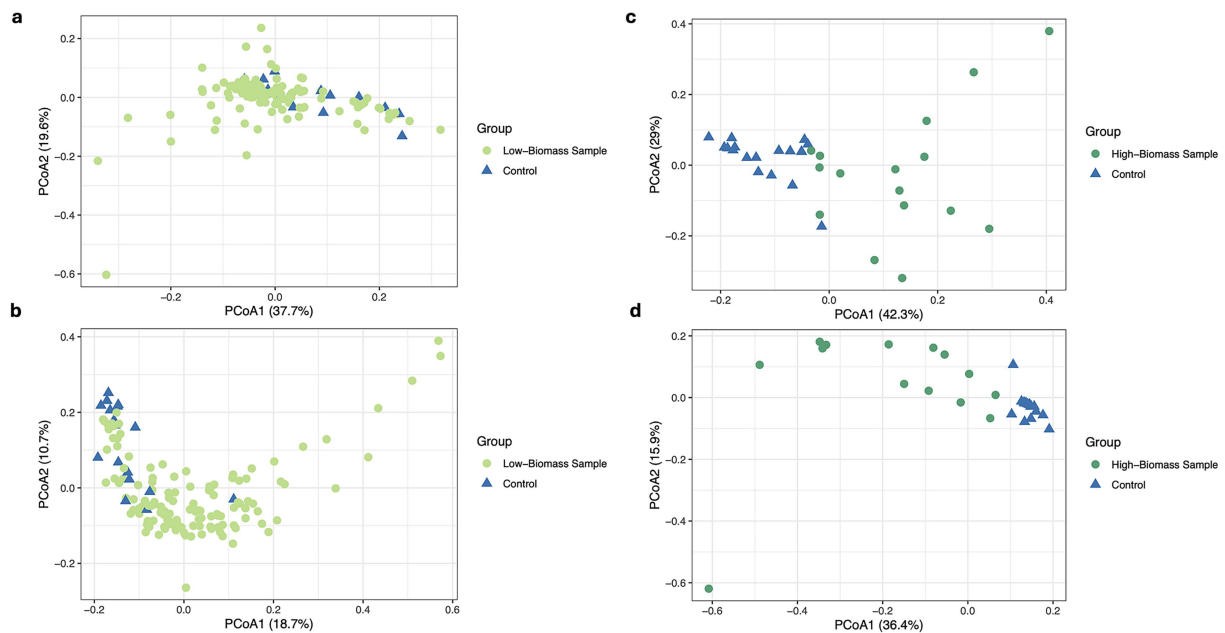


FIGURE 1

Principal coordinates analysis (PCoA) at genus level based on Bray-Curtis distances, showing clinical samples and methodological controls. (a) Low-biomass samples after initial preprocessing (adapter and quality trimming). (b) Low-biomass samples after BBDuk (Bushnell, 2014) read subtraction decontamination strategy. (c) High-biomass samples after initial preprocessing (adapter and quality trimming). (d) High-biomass samples after BBDuk (Bushnell, 2014) read subtraction decontamination.

(<0.01%), indicating low background cross-contamination. However, a few pairs displayed higher misassignment rates (0.2–0.5%), consistent within the previously reported range for index hopping (0.1–2%) (<https://sapac.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing/index-hopping.html>) (Figure 2b).

More specifically, the HCV-spiked sample, which exhibited a strong abundance gradient, low-level HCV reads were detected in other samples, consistent with potential transfer from high-abundance libraries. However, index-sharing analysis does not allow unequivocal determination of read origin, and directionality is inferred based on relative abundance patterns rather than directly observed.

3.3 mNGS diagnostic performance across clinical infection types

3.3.1 Assessment of a chronic infection

We analyzed 17 longitudinal plasma samples from a patient with a chronic HEV infection, viral loads ranged from 35 IU/mL to $\times 2.2 \times 10^6$ IU/mL. Metagenomic sequencing along with background subtraction, successfully detected HEV reads in 15 out of 17 samples. Detection was consistent at viral loads above 10^3 IU/mL, with HEV reads ranging from 82.09 to 7.10×10^5 counts per million (CPM) (Supplementary Table S8). In contrast, samples with viral loads lower than 10^3 IU/mL showed inconsistent or no detection of HEV (Figure 3), despite testing qPCR-positive cases. This suggests that 10^3 IU/mL is the lowest viral load at which detection becomes reproducible and retrieves reliable detection by mNGS. In addition, we observed a strong positive correlation between the log-transformed viral load (IU/mL) and the log-transformed normalized read counts (CPM) obtained by metagenomic sequencing (Pearson's $r = 0.93$, p -value < 0.001), indicating that higher viral loads lead to a

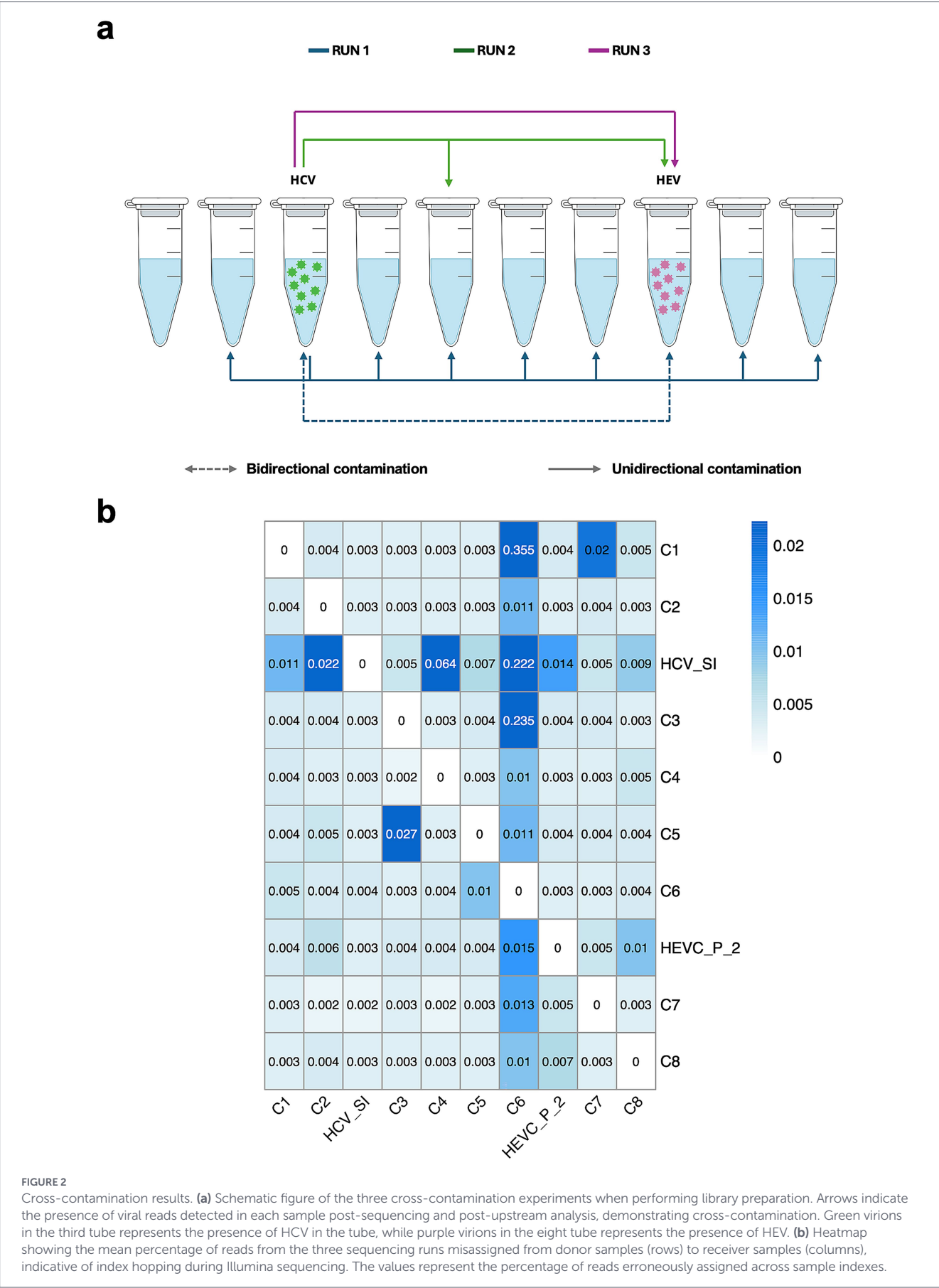
proportionally enhanced detection of the pathogen. This correlation underscores the quantitative potential of metagenomic sequencing in virome studies and highlights detection limits relevant to clinical diagnostics.

In samples with viral loads above 5×10^4 IU/mL, near-complete recovery was achieved, with > 99% genome coverage, CPM values exceeding 170,000, and full-length assemblies generated as single contigs ($N50 = \text{longest contig} = \text{total length} \approx 7,200$ bp) (Supplementary Table S8; Supplementary Figure S8). In contrast, samples with low viral loads (< 10^3 IU/mL), which exhibited inconsistent HEV detection, showed minimal or no genome coverage, CPMs below 1,000, and failed to yield contig assemblies, underscoring the reduced sensitivity of metagenomic sequencing at low input concentrations. Intermediate viral load cases (e.g., HEVC-P-8, HEVC-P-9, HEVC-P-16) produced partial genome coverage (24–75%) and fragmented assemblies composed of multiple shorter contigs with reduced $N50$ values, suggesting a gradual relationship between input viral load and assembly completeness (Supplementary Table S8).

3.3.2 Acute viral infections

We analyzed 14 clinical samples from patients with serologically and RT-qPCR confirmed acute viral infections. These included ten plasma samples (low-biomass) with Cytomegalovirus (CMV), Human polyomavirus 1 (BKV) and Epstein-Barr virus (EBV), with viral loads ranging from 226 IU/mL to 2.87×10^5 IU/mL. In addition, four samples from a patient with acute HEV genotype 1 infection were analyzed: two plasma samples, one amniotic fluid sample (all low biomass), and one fecal sample (high biomass).

mNGS detected its qPCR-confirmed pathogen in eleven out of fourteen samples (Table 3). Consistent with our prior observations, detection became unreliable below viral loads of 10^3 IU/



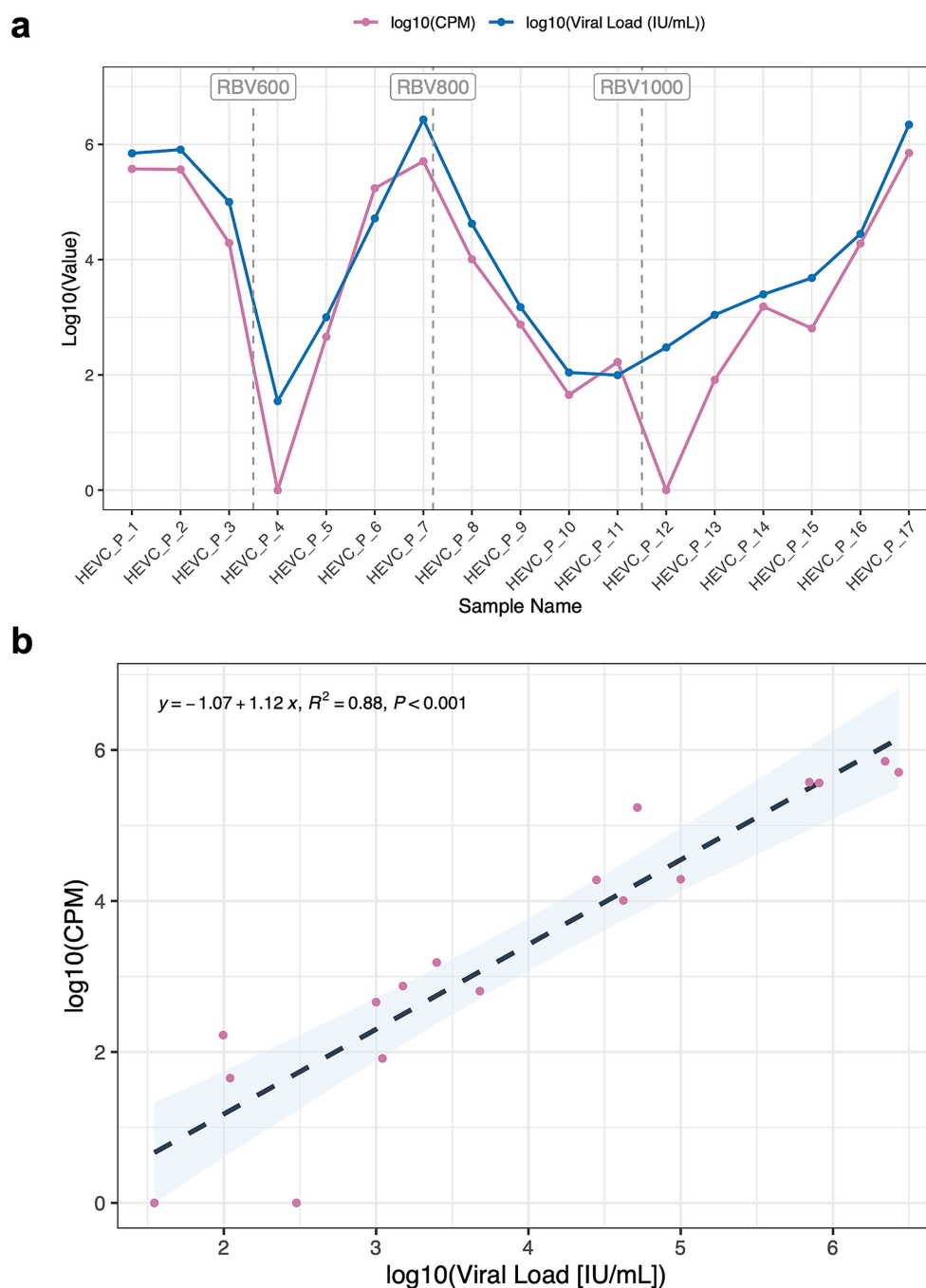


FIGURE 3

Viral load vs metagenomic detection (CPM). **(a)** Line plot showing log-transformed viral load (IU/mL) for hepatitis E virus (HEV) and its corresponding log-transformed viral counts per million. **(b)** Correlation plot between log-transformed viral load and log-transformed normalized read counts (CPM).

mL. Moreover, the log-transformed viral load (IU/mL) strongly correlated with log-transformed CPM values (Pearson's $r = 0.77$, p -value < 0.05), further confirming that higher viral loads correlate with higher read counts in metagenomics, reinforcing the threshold for reliable viral detection in both chronic and acute infection contexts (Supplementary Figure S9). Viral load also correlated positively with genome coverage (Pearson's $r = 0.61$, p -value < 0.05), with coverage spanning from 0% in low-titer samples to up to 100% in two HEV samples (HEVA-P-1, HEVA-F-1). While viral load was the main driver of detection and assembly, some exceptions (PC-P-3) suggest that factors such as read quality, library preparation

biases, or strain divergence also influence assembly and detection success.

Despite successful taxonomic detection, only six of the 14 samples yielded contigs corresponding to the expected pathogen (Table 3; Supplementary Figures S10–S13). Contig recovery ranged from a single 362 bp CMV contig in a plasma sample (viral load of 30,592 IU/mL, sample PC-P-4) to a 7,278 bp HEV contig in the fecal sample HEVA-F-1. Interestingly, additional viral sequences unrelated to the primary infection were identified. In sample PC-P-9, we recovered four contigs attributed to Human betaherpesvirus 6A (N50 = 509 bp, longest contig = 567 bp, total length = 1,940 bp) and three contigs

TABLE 3 Summary of metagenomic sequencing and viral assembly metrics for positive clinical samples with acute infections from Epstein–Barr virus (EBV), cytomegalovirus (CMV) human polyomavirus 1 (BKV), and hepatitis E virus (HEV).

Sample ID	Target virus	Viral load (IU/mL)	CPM	Genome covered (%)	Number of contigs	N50	Longest contig	Total length
PC-P-1	CMV	23,208	13083.15	11.07	6	386	996	2,571
PC-P-2	EBV	25,372	296044.21	74.93	101	667	3,256	63,292
PC-P-3	EBV	2,868,857	2070.93	1.37	NA	NA	NA	NA
PC-P-4	CMV	30,592	2578.88	4.76	1	362	362	362
PC-P-5	CMV	226	43.17	0.08	NA	NA	NA	NA
PC-P-6	EBV	1,430,851	23575.64	16.81	3	996	996	1761
PC-P-7	BKV	4,215	249.07	4.48	NA	NA	NA	NA
PC-P-8	EBV	1,690	0.00	0.00	NA	NA	NA	NA
PC-P-9	EBV	17,832	0.56	0.30	NA	NA	NA	NA
PC-P-10	EBV	4,550	49.65	0.38	NA	NA	NA	NA
HEVA-P-1	HEV	1.01×10 ⁸	163754.15	100	1	7,192	7,192	7,192
HEVA-F-1	HEV	2.1×10 ¹⁰	421568.63	100	1	7,278	7,278	7,278
HEVA-P-2	HEV	120	0.00	0.00	0	NA	NA	NA
HEVA-AF-2	HEV	44	0.00	0.00	0	NA	NA	NA

The table includes Sample ID, target virus identified via qPCR in clinical diagnostics, viral load (IU/mL), counts per million (CPM) of the corresponding virus, percentage of the viral genome covered by reads, number of assembled contigs classified as the target virus, N50, the length of the longest contig, and the length of all viral contigs. NA = not applicable, due to absence of sufficient reads for contig assembly.

corresponding to GB Virus C in the same sample (N50 = 7,767, longest contig = 7,767 bp, total length = 10,020). These findings underscore the potential of mNGS to detect coinfections or latent viral elements beyond the targeted pathogen.

3.3.3 Viral coinfections in respiratory samples

To evaluate mNGS performance in detecting viral coinfections, we analyzed 15 nasopharyngeal swabs and exudates previously tested by RT-qPCR for multiple respiratory viruses, including adenovirus (AdV), bocavirus (BoV), Human coronavirus OC43 (CoV OC43), Human respirovirus 3 (HPIV-3), enterovirus (EV), Human metapneumovirus (MPV), and rhinovirus (RV). The rationale for using respiratory samples to study multiple infections is that the respiratory tract is constantly exposed to environmental microorganisms. Each sample carried coinfections involving two to four viruses, with Ct values spanning from 16.15 to 36.73 for HPIV-3 (Supplementary Table S9). Detection was broadly consistent with RT-qPCR results, showing 84% concordance. Notably, HPIV-3 remained detectable through the technique at Ct values as high as 36.7, suggesting strong sensitivity for this virus. In contrast, ADV and BoV were not consistently above Ct 30. Samples with Ct values < 25 produced high read counts, while detection above Ct > 33 was variable and seemed virus dependent. Additionally, EV, RV and HPIV-3 were identified in samples that had tested negative for these through RT-qPCR (Figure 4). Beyond confirming known coinfections, mNGS revealed additional viral taxa not targeted by routine diagnostics, as seen in the clinical sample PC-P-9. These included Epstein–Barr virus (PC-NFE-3; CPM = 10.08), human betaherpesvirus 6B (PC-NFE-1, CPM = 3.11; PC-NFS-11, CPM = 47.29), human mastadenovirus B, parechovirus A1 (PC-NFS-3, CPM = 4.39), human coronavirus 229E (PC-NFS-2, CPM = 0.91) and influenza C virus (PC-NFS-8, CPM = 12.52), among others detected at very low CPM levels. Notably, sample PC_NFE_1, initially tested positive for CoV OC43

using a commercial multiplex RT-qPCR assay, while mNGS yielded a near-complete genome of human coronavirus HKU1 (longest contig = 22.3 kb).

This suggests possible cross-reactivity or limitations in strain-level resolution in targeted diagnostics and highlights the added value of mNGS for refining or correcting viral classification. Near-complete genomes were recovered for HPIV-3 and RV when RT-qPCR Ct values were < 25, with partial assemblies still achievable up to Ct ≈ 28. At higher Ct, genome recovery was inconsistent, likely reflecting reduced input material (Supplementary Table S9). Some exceptions were observed for small viral genomes (e.g., BoV and RV), which could be assembled even at moderate Ct values. Conversely, fragmented assemblies were obtained for some viruses with low Ct values, such as AdV and EV, likely due to genome complexity, low coverage uniformity, or library preparation bias.

4 Discussion

Our study implemented a comprehensive metagenomic framework to assess contamination control and evaluate the intrinsic sensitivity of shotgun sequencing in real clinical samples. The consistent detection of contaminant genera such as Propionibacterium and Gammaretrovirus across negative controls aligns with previous reports (Gu et al., 2019; Jurasz et al., 2021; Lou et al., 2023; Zinter et al., 2019; Salter et al., 2014) on the “kitome,” reinforcing the need for rigorous contamination filtering to mitigate their dominating effect, specifically in low-biomass samples, where these can dominate over genuine clinical signals (Zinter et al., 2019).

These findings support the feasibility of building cross-laboratory contaminant reference sets to complement local filtering. Importantly, this contaminant watchlist is taxon-based: the recurrent detection of reads assigned to a given genus in negative controls does not

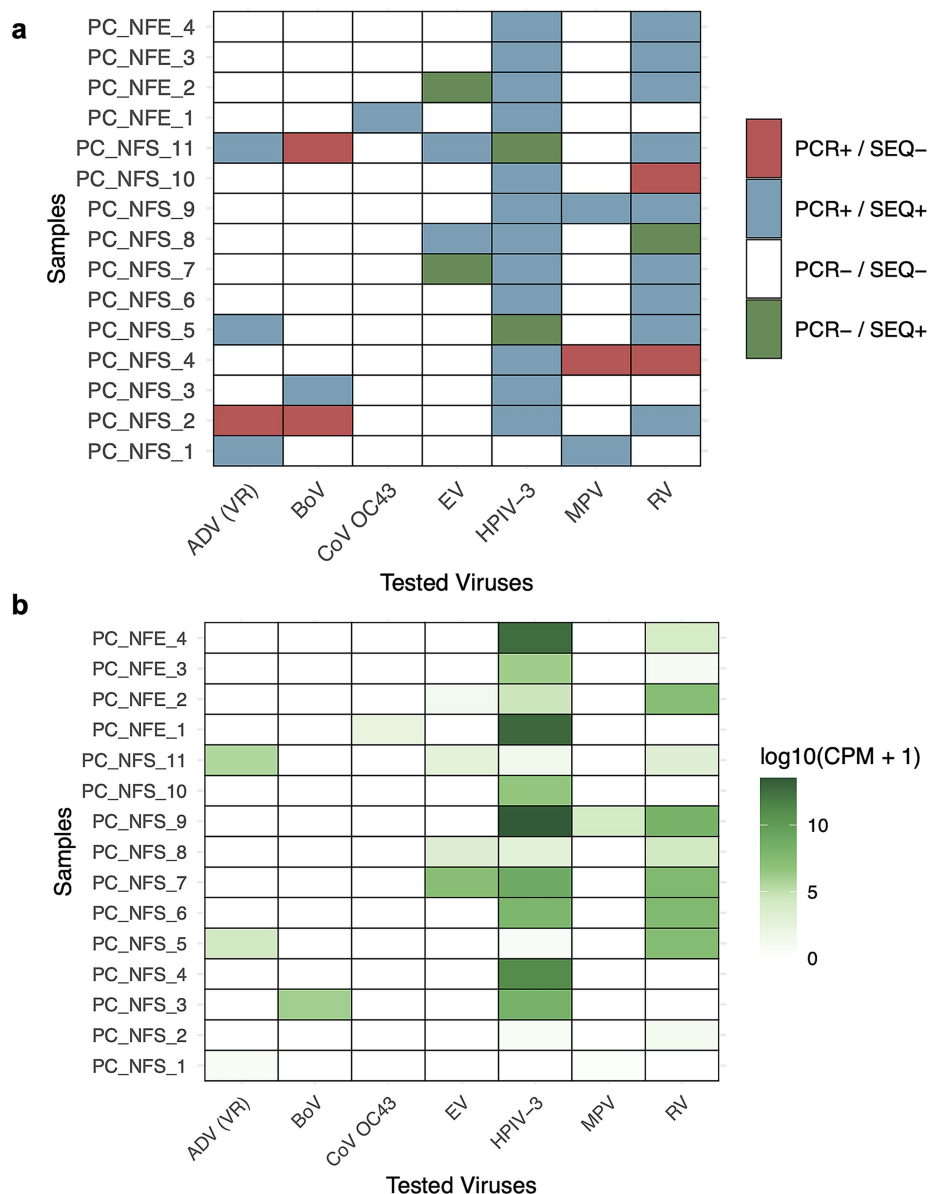


FIGURE 4

Detection of respiratory viruses in coinfections. **(a)** Heatmap summarizing the detection status of tested viruses across all samples using qPCR and sequencing. Each tile represents a sample-virus pair and can be found color-coded by detection outcome. **(b)** Heatmap showing log-transformed viral read counts (CPM) of virus-specific sequencing reads across samples. CPM was calculated as the number of virus-specific sequencing reads normalized to the total number of non-host reads per sample. Each tile represents a sample-virus pair and is color-coded by normalized viral sequencing abundance. The tested viruses included adenovirus [ADV (VR)], bocavirus (BoV), human coronavirus OC43 (CoV OC43), enterovirus (EV), human respirovirus 3 (HPIV-3), human metapneumovirus (MPV), and rhinovirus (RV).

necessarily imply that the entire organism is intrinsically a contaminant, but rather that low-level signals from this taxon are consistently observed across runs. The source of contaminant sequences in metagenomic datasets often remains unclear, potentially arising from reagents, laboratory environments, sample collection and personnel handling during extraction and library preparation, emphasizing the need for standardized contamination management practices. This could be done by implementing rigorous negative controls, developing lab-specific contaminant watchlists, and incorporating computational subtraction strategies (Hornung et al., 2019; Porter et al., 2021), such as decontam (Davis et al., 2018) or metathresholds (<https://github.com/sarah-buddle/metathresholds>) to distinguish true biological signals from background noise. However, distinguishing true signals

from contamination remains particularly challenging for taxa that are biologically plausible in the sample context (e.g., skin or respiratory-associated microbes), underscoring the need to combine wet- and dry-lab decontamination strategies. Therefore, the contaminant watchlist we have generated is not intended as an exclusion list, but as an interpretative framework to contextualize genera identified in samples. Clinical interpretation should integrate both abundance metrics, coverage distribution, and clinical context before attributing pathogenic relevance.

Moreover, we acknowledge that the specific composition of the contaminant watchlist is inherently dependent on the local laboratory environment (extraction kits, library preparation methods, operational workflows...). Therefore, the exact taxa identified in this study

may not be universally generalizable across institutions, although some contaminants have already been reported in the literature, as stated in the results.

Our bioinformatics approach, which combined BBDuk (Bushnell, 2014) read subtraction with prevalence-based filtering using decontam (Davis et al., 2018), enhanced community differentiation beyond what either method achieved alone, while increasing within-group dispersion. This likely arises from the compounded effects of aggressive contaminant filtering and biological heterogeneity in low-biomass samples. This is not necessarily undesirable, it may signify genuine biological heterogeneity, consistent with known principles of microbiome variation in low-biomass or stressed environments (Davis et al., 2018; Austin and Korem, 2024). Although the analytical workflow used in this study is based on previously described bioinformatic approaches, our objective was to evaluate its performance and practical applicability in real-world clinical samples, including plasma and respiratory specimens with varying viral loads. Our results provide an integrated assessment of its ability to detect and characterize viral sequences in challenging low-biomass samples, highlighting both its diagnostic potential and technical limitations. This type of systematic evaluation in clinical specimens is essential to support the implementation of metagenomic sequencing in routine diagnostic settings, where factors such as host background, sequencing depth, and analytical filtering can significantly influence performance.

In high-biomass samples, contaminants contribute less compared to true biological content, reducing the impact on downstream analyses. As a result, such samples may not require the same level of decontamination stringency as low-biomass samples although contamination-aware workflows and negative controls remain essential (Jurasz et al., 2021). Moreover, persistent contamination underscores a major clinical risk, potentially causing false positives, inappropriate treatment, or missed infections.

While environmental and reagent-based contamination has been widely studied cross-sample contamination remains comparatively unexplored (Jurasz et al., 2021; Lou et al., 2023; Zinter et al., 2019; Salter et al., 2014). To address this, we conducted a controlled sequencing run using spiked and unspiked negative controls. Despite stringent procedures, read carryover was observed, showing that cross-sample contamination can compromise interpretation even under controlled conditions, especially in low-biomass samples. This highlights the need for careful experimental designs that include multiple types of negative controls (Fierer et al., 2025) (sampling, extraction and library). Dual-index misassignment further illustrates how technical artifacts can mimic coinfections or suggest false pathogens. Internal spike-in controls may help establish run-specific positivity thresholds but replicating this contamination assay across additional independent sequencing runs are needed to confirm reproducibility. Future advances in library preparation, barcoding, and QC metrics could further reduce index bleed-through and enhance mNGS reliability.

In viral diagnostics, metagenomic sequencing offers an untargeted approach for pathogen detection and genome characterization (Gu et al., 2019). However, its sensitivity remains a key limitation compared to targeted molecular assays such as qPCR (Ibañez-Llagoña et al., 2023), being highly dependent on input viral load. To establish clinically actionable sensitivity benchmarks, we analyzed samples from three representative infection contexts: chronic infections, acute infections, and respiratory co-infections, ensuring thresholds are directly applicable to real-world diagnostic practices. For the chronic infection context, we analyzed plasma samples from a patient with chronic HEV

infection undergoing ribavirin treatment. Ribavirin, a mutagenic nucleoside analog, complicates genome recovery as it increases intra-host diversity and the burden of low-frequency mutations (Colomer-Castell et al., 2023). These changes hinder consensus accuracy and full-genome assembly, which complicates genome characterization when relying on specific primers, and for which metagenomics could provide an advantage. Our findings show that metagenomic sequencing can detect and recover HEV genomes in chronic infections, with reliable detection at 10^3 IU/mL and nearly complete genome recovery only achievable when viral loads exceeded $\sim 5 \times 10^4$ IU/mL. Thus, despite its value for unbiased pathogen discovery, metagenomics cannot replace qPCR for pathogen detection yet and integrating viral load data remains essential for interpreting results.

Extending the analysis to acute infections, we applied mNGS to thirteen low-biomass plasma samples and one high-biomass fecal sample from patients with acute infections. Despite the challenges of working with low-biomass samples, metagenomics detected the target virus in eleven out of the thirteen plasma samples, demonstrating its potential utility for untargeted diagnostics. As observed in the HEV chronic infection series, detection strongly correlated with viral load and samples with viral loads below 10^3 IU/mL failed to yield detectable metagenomic evidence of infection, reinforcing the existence of a practical sensitivity threshold for this approach (Buddle et al., 2024). Contig-level identification was only achieved in six of the samples, highlighting the importance read-level classification.

Near-complete HEV genomes were recovered from both acute infection samples with high viral loads ($\sim 10^8$ – 10^{10} IU/mL; HEVA-P-1 and HEVA-F-1), illustrating how elevated viral titers substantially enhance genome recovery. It is important to note that HEV's small genome (~ 7.2 kb) likely facilitates its amplification and assembly in metagenomic workflows (Nayfach et al., 2020; Smits et al., 2015; Smits et al., 2014). Furthermore, the presence of HEV in plasma as intact viral particles contributes to deep and uniform coverage, improving the chances of assembling complete or near-complete genomes (Shahini et al., 2024; Manso et al., 2017).

In contrast, CMV, EBV and BKV are predominantly present in plasma as fragmented cell-free DNA (Parikh and Anderson, 2025; Peddu et al., 2020; Lam et al., 2018), making full-genome assembly more challenging, yet metagenomics is able to detect them (Sam et al., 2021). Importantly, this approach also revealed coinfections, reinforcing the power of mNGS as a diagnosis tool and highlighting its potential to uncover infections often missed by conventional diagnostics. Metagenomic sequencing was further evaluated in nasopharyngeal swabs and exudate samples from patients with RT-qPCR-confirmed coinfections involving two to four respiratory viruses. Metagenomic detection was broadly concordant with RT-qPCR.

However, while molecular diagnostics offer high sensitivity and specificity, they are inherently limited by their targeted nature and may miss unexpected or emerging viruses. Moreover, its implementation in routine clinical practice remains constrained by substantial financial costs, the need for specialized laboratory infrastructure, and advanced bioinformatic expertise. In addition, standardized analytical pipelines, quality control frameworks and regulatory accreditation for clinical interpretation are not yet universally established. In contrast, metagenomics is capable of simultaneously detecting known pathogens, uncovering additional viral agents not included in diagnostic panels, and even resolving strain-level identities. For instance, the recovery of a near-complete genome of human coronavirus HKU1 in a sample initially diagnosed as CoV OC43 highlights the added value

of mNGS in refining viral classification and detecting closely related species.

Nevertheless, some viruses such as ADV and EV, were more difficult to assemble, possibly due to greater genome complexity or uneven read distribution. Beyond confirming known infections, mNGS also detected additional viruses such as EBV or human herpesvirus 6B, in addition to viruses which had tested negative through RT-qPCR (HPIV-3, RV and EV). These findings highlight metagenomics' ability to uncover unexpected coinfections, raising questions about their clinical relevance, especially in immunocompromised hosts or in cases where latent or reactivating viruses may contribute to disease. Moreover, we were able to recover almost full genomes for HPIV-3, RV and human coronavirus HKU1, possibly due to smaller genome sizes (Nayfach et al., 2020; Smits et al., 2015; Smits et al., 2014) and the nature of NP swabs, which typically exhibit lower levels of host nucleic acid background (Wylie et al., 2012), thereby improving signal-to-noise ratios.

Despite the effectiveness of our combined decontamination strategy, distinguishing true low-abundance signals from residual contamination remains challenging. Including negative controls that match each sample could improve baselines and interpretation of borderline detections. Additionally, while Kraken2 (Wood et al., 2019) provides rapid and computationally efficient taxonomic classification, its reliance on k-mers can introduce misclassifications due to conserved sequences which may be shared across phylogenetically diverse organisms. Consequently, low-level taxonomic assignments to unexpected environmental taxa may arise. Such signals may therefore reflect systematic classification background inherent to database structure and k-mer matching strategies, rather than true biological presence. Importantly, our objective was not to infer the biological origin of these taxa, but to identify recurrent background assignments observed in negative controls and across sample types. Future studies may benefit from combining multiple classifiers or applying alignment-based confirmation to improve accuracy and taxonomic resolution.

The key contribution for this study is a pragmatic framework for establishing and applying a contamination watchlist in clinical shotgun metagenomic workflows, particularly for low-biomass samples. The proposed approach consists of: (1) generating multiple methodological negative controls, including both extraction and library controls processed alongside clinical samples in multiple sequencing runs; (2) sequencing these controls using the same laboratory and bioinformatic pipeline; (3) identifying taxa that are recurrently detected across controls; and (4) flagging these taxa as potential contaminants candidates during interpretation of clinical datasets. Rather than attempting to eliminate all background taxa, this framework aims to contextualize detections to reduce overinterpretation. Moreover, methodological controls should be added in all forthcoming sequencing runs as contamination monitoring is dynamic due to batch-to-batch variations in reagents, laboratory conditions, and handling practices that can introduce new contaminants over time. We do not propose a fixed number of controls as universally sufficient. Rather, laboratories should generate enough controls to allow recurrent background taxa to stabilize. The number required will depend on workflow complexity, sample types to be analyzed, and reagent variability.

The benefit of this approach is interpretative clarity: by pre-characterizing recurrent background taxa, laboratories can substantially reduce the number of genera requiring manual clinical review and focus attention on biologically plausible, sample-consistent organisms, which provides a structured interpretative layer to improve reporting consistency and transparency in clinical metagenomic diagnostics.

In summary, our study systematically evaluated contamination in viral metagenomics, highlighting the necessity of rigorous decontamination workflows, particularly in low-biomass samples frequently used in clinical settings. By integrating methodological controls, computational filtering, and lab-specific contaminant watchlists, we present a reproducible framework that enhances data interpretation and genome recovery in clinical metagenomics. Importantly, our analyses across different infections establish context-specific benchmarks, enabling laboratories to interpret metagenomics results with greater confidence. These findings underscore the continued need for contamination-aware pipelines and standardized controls in clinical metagenomics and suggest that, while mNGS holds substantial promise for broad-spectrum pathogen detection, its accuracy hinges on robust experimental design and data curation. Moving forward, the integration of quantitative measures (e.g., viral load), improved decontamination tools, and cross-laboratory contaminant reference sets will be critical to enhance diagnostic reliability and biological insight.

Data availability statement

The datasets generated and analyzed in this study are publicly available in the NCBI SRA repository (Sequence Read Archive) under the BioProject accession number PRJNA1424541. All scripts used for the data analysis will be made available upon request by contacting the corresponding author via email.

Ethics statement

The studies involving humans were approved by Vall d'Hebron Barcelona Hospital Campus Ethics Committee. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

MI-L: Software, Formal analysis, Writing – review & editing, Writing – original draft, Methodology, Visualization, Investigation, Conceptualization. SC-C: Visualization, Investigation, Writing – original draft, Methodology. CC: Investigation, Writing – review & editing, Methodology, Visualization. ÁG-C: Investigation, Writing – review & editing, Methodology. AL: Project administration, Writing – review & editing, Methodology, Investigation, Resources, Visualization. JG-L: Resources, Project administration, Validation, Visualization, Conceptualization, Writing – review & editing, Investigation. DS-T: Methodology, Investigation, Resources, Writing – review & editing. AR-S: Writing – review & editing, Methodology. CA: Methodology, Resources, Writing – review & editing. JE: Resources, Methodology, Writing – review & editing. PN: Resources, Methodology, Writing – review & editing. RF: Methodology, Resources, Writing – review & editing. MC: Writing – review & editing, Conceptualization, Resources. DT: Writing – review & editing, Conceptualization, Resources. JG: Resources, Conceptualization, Writing – review &

editing. MR-B: Resources, Formal analysis, Validation, Writing – review & editing. JR-C: Formal analysis, Validation, Resources, Writing – review & editing. AR: Writing – review & editing, Resources, Conceptualization. EB: Resources, Writing – review & editing, Validation, Formal analysis. MB: Formal analysis, Validation, Resources, Writing – review & editing. MG: Writing – review & editing, Conceptualization, Funding acquisition. AA: Funding acquisition, Conceptualization, Writing – review & editing. AC: Funding acquisition, Writing – review & editing, Conceptualization. RJ-M: Funding acquisition, Writing – review & editing, Conceptualization. JQ: Project administration, Validation, Methodology, Supervision, Funding acquisition, Writing – original draft.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This study was partially supported by Grants PI22/00258, PI22/00023, and PI23/01065 funded by Instituto de Salud Carlos III (ISCIII), Grant PMPER24/00018 and PID2021-126447OB-I00 under the Ministry of Science, Innovation and Universities of Spain and cofounded by the European Union; and Consorcio Centro de Investigación Biomédica en Red CIBERINFEC and CIBEREHD. MI-L was supported by a fellowship from the “La Caixa” Foundation (ID 100010434) with code “LCF/BQ/DR23/12000020.” SC-C received support from the Spanish Ministry of Education grant FPU21/04150. CC received a predoctoral fellowship from Vall d’Hebron Institute of Research (VHIR).

Acknowledgments

We would like to acknowledge Roche Diagnostics S. L. U. for supporting the reagents through the Kapamet project and financing the cost of publishing this manuscript.

References

- Asplund, M., Kjartansdóttir, K. R., Møllerup, S., Vinner, L., Fridholm, H., Herrera, J. A. R., et al. (2019). Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. *Clin. Microbiol. Infect.* 25, 1277–1285. doi: 10.1016/j.cmi.2019.04.028
- Austin, G. I., and Korem, T. (2024). Planning and analyzing a low-biomass microbiome study: a data analysis perspective. *J. Infect. Dis.* 233, 76–86. doi: 10.1093/infdis/jiae378
- Babraham Bioinformatics. FastQC A Quality Control tool for High Throughput Sequence Data. (2010). Available online at: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114. doi: 10.1093/BIOINFORMATICS/BTU170
- Buddle, S., Forrest, L., Akisuyi, N., Martin Bernal, L. M., Brooks, T., Venturini, C., et al. (2024). Evaluating metagenomics and targeted approaches for diagnosis and surveillance of viruses. *Genome Med.* 16:111. doi: 10.1186/S13073-024-01380-X
- Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. (2014).
- Cano, A., Capdevila, M., Puerta, R., Arranz, J., Montreal, L., de Rojas, I., et al. (2024). Clinical value of plasma pTau181 to predict Alzheimer’s disease pathology in a large real-world cohort of a memory clinic. *EBioMedicine* 108:105345. doi: 10.1016/j.ebiom.2024.105345
- Chiu, C. Y., and Miller, S. A. (2019). Clinical metagenomics. *Nat. Rev. Genet.* 20, 341–355. doi: 10.1038/s41576-019-0113-7

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmicb.2026.1786985/full#supplementary-material>

- Chrzastek, K., Tennakoon, C., Bialy, D., Freimanis, G., Flannery, J., and Shelton, H. (2022). A random priming amplification method for whole genome sequencing of SARS-CoV-2 virus. *BMC Genomics* 23, 1–20. doi: 10.1186/s12864-022-08563-z
- Colomer-Castell, S., Gregori, J., Garcia-Cehic, D., Riveiro-Barciela, M., Buti, M., Rando-Segura, A., et al. (2023). In-host HEV Quasispecies evolution shows the limits of mutagenic antiviral treatments. *Int. J. Mol. Sci.* 24:17185. doi: 10.3390/IJMS242417185
- Comas, I., Cancino-Muñoz, I., Mariner-Llicer, C., Goig, G. A., Ruiz-Hueso, P., Francés-Cuesta, C., et al. (2020). Use of next generation sequencing technologies for the diagnosis and epidemiology of infectious diseases. *Enfermedades infecciosas y microbiología clínica (English ed)*. 38, 32–38. doi: 10.1016/j.eimc.2020.02.006
- Corless, C. E., Guiver, M., Borrow, R., Edwards-Jones, V., Kaczmarek, E. B., and Fox, A. J. (2000). Contamination and sensitivity issues with a real-time universal 16S rRNA PCR. *J. Clin. Microbiol.* 38:1747. doi: 10.1128/JCM.38.5.1747-1752.2000
- Davis, N. M., Proctor DiM, D., Holmes, S. P., Relman, D. A., and Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 6:226. doi: 10.1186/S40168-018-0605-2
- Edridge, A. W. D., Deijs, M., Van Zeggeren, I. E., Kinsella, C. M., Jebbink, M. F., Bakker, M., et al. (2019). Viral metagenomics on cerebrospinal fluid. *Genes* 10:332. doi: 10.3390/GENES10050332
- Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. doi: 10.1093/BIOINFORMATICS/BTW354

- Fierer, N., Leung, P. M., Lappan, R., Eisenhofer, R., Ricci, F., Holland, S. I., et al. (2025). Guidelines for preventing and reporting contamination in low-biomass microbiome studies. *Nat. Microbiol.* 10, 1570–1580. doi: 10.1038/s41564-025-02035-2
- Gu, W., Miller, S., and Chiu, C. Y. (2019). Clinical metagenomic next-generation sequencing for pathogen detection. *Annu. Rev. Pathol.* 14, 319–338. doi: 10.1146/annurev-pathmechdis-012418-012751
- Hornung, B. V. H., Zwiittink, R. D., and Kuijper, E. J. (2019). Issues and current standards of controls in microbiome research. *FEMS Microbiol. Ecol.* 95:45. doi: 10.1093/FEMSEC/FIZ045
- Ibañez-Lligoña, M., Colomer-Castell, S., González-Sánchez, A., Gregori, J., Campos, C., García-Cehic, D., et al. (2023). Bioinformatic tools for NGS-based metagenomics to improve the clinical diagnosis of emerging, re-emerging and new viruses. *Viruses* 15:587. doi: 10.3390/v15020587
- Iyer, R., and Damania, A. (2020). Shotgun metagenomics of indigenous bacteria collected from the banks of the San Jacinto River for biodegradation of aromatic waste. *FEMS Microbiol. Lett.* 367:133. doi: 10.1093/FEMSLE/FNAA133
- Jurasz, H., Pawłowski, T., and Perlejewski, K. (2021). Contamination issue in viral metagenomics: problems, solutions, and clinical perspectives. *Front. Microbiol.* 12:12. doi: 10.3389/FMICB.2021.745076
- Lam, W. K. J., Jiang, P., Chan, K. C. A., Cheng, S. H., Zhang, H., Peng, W., et al. (2018). Sequencing-based counting and size profiling of plasma Epstein-Barr virus DNA enhance population screening of nasopharyngeal carcinoma. *Proc. Natl. Acad. Sci. USA* 115, E5115–E5124. doi: 10.1073/PNAS.1804184115
- Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9:357–359. doi: 10.1038/nmeth.1923
- Li, D., Liu, C. M., Luo, R., Sadakane, K., and Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/BIOINFORMATICS/BTV033
- Lou, Y. C., Hoff, J., Olm, M. R., West-Roberts, J., Diamond, S., Firek, B. A., et al. (2023). Using strain-resolved analysis to identify contamination in metagenomics data. *Microbiome* 11, 1–14. doi: 10.1186/s40168-023-01477-2
- Manso, C. F., Bibby, D. F., and Mbisa, J. L. (2017). Efficient and unbiased metagenomic recovery of RNA virus genomes from human plasma samples. *Sci. Rep.* 7:4173. doi: 10.1038/S41598-017-02239-5
- Miao, Q., Ma, Y., Wang, Q., Pan, J., Zhang, Y., Jin, W., et al. (2018). Microbiological diagnostic performance of metagenomic next-generation sequencing when applied to clinical practice. *Clin. Infect. Dis.* 67, S231–S240. doi: 10.1093/CID/CY693
- Miller, S., Naccache, S. N., Samayoa, E., Messacar, K., Arevalo, S., Federman, S., et al. (2019). Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome Res.* 29, 831–842. doi: 10.1101/GR.238170.118/-/DC1
- Minich, J. J., Sanders, J. G., Amir, A., Humphrey, G., Gilbert, J. A., and Knight, R. (2019). Quantifying and understanding well-to-well contamination in microbiome research. *mSystems* 4:e00186–19. doi: 10.1128/MSYSTEMS.00186-19
- Nayfach, S., Camargo, A. P., Schulz, F., Eloë-Fadros, E., Roux, S., and Kyrpides, N. C. (2020). Checkv assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* 39, 578–585. doi: 10.1038/s41587-020-00774-7
- Oksanen, J., Simpson, G. J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., et al. (2025). Community Ecology Package [R Package version 2.7–2]. (2025)
- Olomu, I. N., Pena-Cortes, L. C., Long, R. A., Long, R. A., Vyas, A., Krichevskiy, O., et al. (2020). Elimination of “kitome” and “splashome” contamination results in lack of detection of a unique placental microbiome. *BMC Microbiol.* 20:157. doi: 10.1186/S12866-020-01839-Y
- Parikh, B. A., and Anderson, N. W. (2025). The impact of FDA-cleared molecular solutions for BK polyomavirus quantitation. *J. Clin. Microbiol.* 63:e00348–24. doi: 10.1128/JCM.00348-24
- Parras-Moltó, M., Rodríguez-Galet, A., Suárez-Rodríguez, P., and López-Bueno, A. (2018). Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses. *Microbiome* 6, 1–18. doi: 10.1186/S40168-018-0507-3
- Peddu, V., Bradley, B. T., Casto, A. M., Shree, R., Colbert, B. G., Xie, H., et al. (2020). High-resolution profiling of human cytomegalovirus cell-free DNA in human plasma highlights its exceptionally fragmented nature. *Sci. Rep.* 10:3734. doi: 10.1038/S41598-020-60655-6
- Piro, V. C., and Renard, B. Y. (2022). Contamination detection and microbiome exploration with GRIMER. *Gigascience* 12:12. doi: 10.1093/GIGASCIENCE/GIAD017
- Porter, A. F., Cobbin, J., Li, C. X., Eden, J. S., and Holmes, E. C. (2021). Metagenomic identification of viral sequences in laboratory reagents. *Viruses* 13:2122. doi: 10.3390/v13112122
- Quer, J., Colomer-Castell, S., Campos, C., Andrés, C., Piñana, M., Cortese, M. F., et al. (2022). Next-generation sequencing for confronting virus pandemics. *Viruses* 14:600. doi: 10.3390/v14030600
- R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; (2022). Available online at: <https://www.R-project.org/> (Accessed June 8, 2025).
- Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12:87. doi: 10.1186/S12915-014-0087-Z
- Sam, S. S., Rogers, R., Gillani, F. S., Tsongalis, G. J., Kraft, C. S., and Caliendo, A. M. (2021). Evaluation of a next-generation sequencing metagenomics assay to detect and quantify DNA viruses in plasma from transplant recipients. *J. Mol. Diagn.* 23:719. doi: 10.1016/J.JMOLDX.2021.02.008
- Schloss, P. D., and Handelsman, J. (2005). Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol.* 6:229. doi: 10.1186/GB-2005-6-8-229
- Shahini, E., Argentiero, A., Andriano, A., Losito, F., Maida, M., Facciorusso, A., et al. (2024). Hepatitis E virus: what more do we need to know? *Medicina* 60:998. doi: 10.3390/MEDICINA60060998
- Smits, S. L., Bodewes, R., Ruiz-González, A., Baumgärtner, W., Koopmans, M. P., Osterhaus, A. D. M. E., et al. (2015). Recovering full-length viral genomes from metagenomes. *Front. Microbiol.* 6:165597. doi: 10.3389/fmicb.2014.00714
- Smits, S. L., Bodewes, R., Ruiz-Gonzalez, A., Baumgärtner, W., Koopmans, M. P., Osterhaus, A. D. M. E., et al. (2014). Assembly of viral genomes from metagenomes. *Front. Microbiol.* 5:118886. doi: 10.3389/FMICB.2014.00714/BIBTEX
- Su, Q., Wong, O. W. H., Lu, W., Wan, Y., Zhang, L., Xu, W., et al. (2024). Multikindom and functional gut microbiota markers for autism spectrum disorder. *Nat. Microbiol.* 9, 2344–2355. doi: 10.1038/s41564-024-01739-1
- von Fricken, M. E., Melendrez, M. C., Linton, Y. M., and Takhampunya, R. (2023). Metagenomics for epidemiological surveillance in one health. *Front. Microbiol.* 14:1191946. doi: 10.3389/FMICB.2023.1191946
- Wickham, H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; (2016). Available online at: <https://ggplot2.tidyverse.org> (Accessed June 8, 2025).
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., et al. (2019). Welcome to the tidyverse. *J. Open Source Softw.* 4:1686. doi: 10.21105/joss.01686
- Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome Biol.* 20, 1–13. doi: 10.1186/s13059-019-1891-0
- Wylie, K. M., Mihindukulasuriya, K. A., Sodergren, E., Weinstock, G. M., and Storch, G. A. (2012). Sequence analysis of the human Virome in febrile and afebrile children. *PLoS One* 7:e27735. doi: 10.1371/JOURNAL.PONE.0027735
- Xie, Z., Canalda-Baltrons, A., d’Enfert, C., and Manichanh, C. (2023). Shotgun metagenomics reveals interkingdom association between intestinal bacteria and fungi involving competition for nutrients. *Microbiome* 11:275. doi: 10.1186/s40168-023-01693-w
- Zhao, Y., Zhang, W., and Zhang, X. (2024). Application of metagenomic next-generation sequencing in the diagnosis of infectious diseases. *Front. Cell. Infect. Microbiol.* 14:1458316. doi: 10.3389/fcimb.2024.1458316
- Zinter, M. S., Mayday, M. Y., Ryckman, K. K., Jelliffe-Pawlowski, L. L., and Derisi, J. L. (2019). Towards precision quantification of contamination in metagenomic sequencing experiments. *Microbiome* 7, 1–5. doi: 10.1186/S40168-019-0678-6/FIGURES/3