frontiers in Microbiology

REVIEW
published: 20 October 2021
doi: 10.3389/fmicb.2021.745076

Contamination Issue in Viral Metagenomics: Problems, Solutions, and Clinical Perspectives

*Henryk Jurasz[1], Tomasz Pawłowski[2] and Karol Perlejewski[1]**

[1] Department of Immunopathology of Infectious and Parasitic Diseases, Medical University of Warsaw, Warsaw, Poland,
[2] Division of Psychotherapy and Psychosomatic Medicine, Department of Psychiatry, Wrocław Medical University, Wrocław, Poland

We describe the most common internal and external sources and types of contamination encountered in viral metagenomic studies and discuss their negative impact on sequencing results, particularly for low-biomass samples and clinical applications. We also propose some basic recommendations for reducing the background noise in viral shotgun metagenomic (SM) studies, which would limit the bias introduced by various classes of contaminants. Regardless of the specific viral SM protocol, contamination cannot be totally avoided; in particular, the issue of reagent contamination should always be addressed with high priority. There is an urgent need for the development and validation of standards for viral metagenomic studies especially if viral SM protocols will be more widely applied in diagnostics.

Keywords: viral metagenomics, virome, contamination, low-biomass sample, virus

OPEN ACCESS

Edited by:
Matthias Hess, University of California, Davis, United States

Reviewed by:
Simon Roux, Joint Genome Institute, Lawrence Berkeley National Laboratory, United States
Karthik Anantharaman, University of Wisconsin-Madison, United States

****Correspondence:***
Karol Perlejewski
kperlejewski@wum.edu.pl

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *22 July 2021*
Accepted: *17 September 2021*
Published: *20 October 2021*

Citation:
Jurasz H, Pawłowski T and Perlejewski K (2021) Contamination Issue in Viral Metagenomics: Problems, Solutions, and Clinical Perspectives. Front. Microbiol. 12:745076. doi: 10.3389/fmicb.2021.745076

INTRODUCTION

Next-generation sequencing (NGS) techniques combined with the development of computational tools led to an explosion of metagenomic studies in the past decade (Chiu and Miller, 2019; Lewandowski et al., 2019). Metagenomics is defined as direct analysis of the whole microbial communities based on DNA/RNA extracted from clinical or environmental samples (Huson and Mitra, 2012). Such analysis allows for the detection of known and unknown microorganisms and provides insights into the pathogen–host interactions, epidemiology, ecology, and evolution of organisms found across various ecosystems (Forbes et al., 2017; Chiu and Miller, 2019). Although microbial research remains dominated by bacterial 16S rRNA gene sequencing studies, new techniques were also used for viral analysis (Ladner et al., 2014; Moustafa et al., 2017; Kufner et al., 2019). Shotgun metagenomics (SM) is currently the most widely used technique to analyze viral DNA and RNA in a given environment (Conceicao-Neto et al., 2015; Forbes et al., 2017) and was successfully introduced into clinical practice to support diagnosis of systemic infections and occasionally identified a number of novel viral species (Palacios et al., 2008; Foulongne et al., 2011; Lipowski et al., 2017).

While SM is being used to characterize the virome using various workflows, it still faces numerous challenges, including the decision regarding best extraction and sequencing methods, the need for host genomic background depletion, the necessity of access to computational resources and highly specialized bioinformaticists, and providing relevant clinical data fast enough to be of clinical value (Schlaberg et al., 2017; Boers et al., 2019). Overall, SM approach has allowed

for comprehensive surveys of never-before-seen viral communities (Moreno-Gallego et al., 2019; Waldvogel-Abramowski et al., 2019; Perlejewski et al., 2020b). However, SM also detects external contaminant nucleic acids and cross-contaminations, which can affect the interpretation of microbiome data (Xu et al., 2013; Laurence et al., 2014). So far, the issue of contamination in microbial sequencing studies was mostly discussed in regard to amplicon target sequencing (ATM); (16S rRNA gene sequencing) and was largely focused on bacterial bias (Karstens et al., 2019). Such contamination effects are common, as several studies have found contaminant microbial DNA in laboratory reagents and laboratory surfaces (Salter et al., 2014; Eisenhofer et al., 2019; Stinson et al., 2019). While several groups have also reported on the presence of genomic contaminants in viral SM data, there are no established criteria for examination and/or reporting of contamination in virome-focused studies (Moustafa et al., 2017; Zolfo et al., 2019; Perlejewski et al., 2020a). The current review emphasizes the impact of contaminants on viral studies, especially when using low-biomass samples, and proposes recommendations to minimize its effect.

SOURCES OF CONTAMINATION IN MICROBIOME STUDIES

Different types of samples and SM protocols affect the composition of genetic background found in viral metagenomics. Therefore, contaminants may be represented by external host/human or bacterial DNA, as well as sequencing reads aligned to genomes of a non-sample viral, fungal, protozoal, or even plant species (Perlejewski et al., 2015; Moustafa et al., 2017; Asplund et al., 2019). Specific contaminants are often not even reported in viral metagenomic studies as most viral SM research is focused only on viral hits, rarely aligning NGS reads to genomes other than host and viral. There are two major types of contaminants in viral SM studies: external or internal contamination (**Figure 1**; Davis et al., 2018; Eisenhofer et al., 2019).

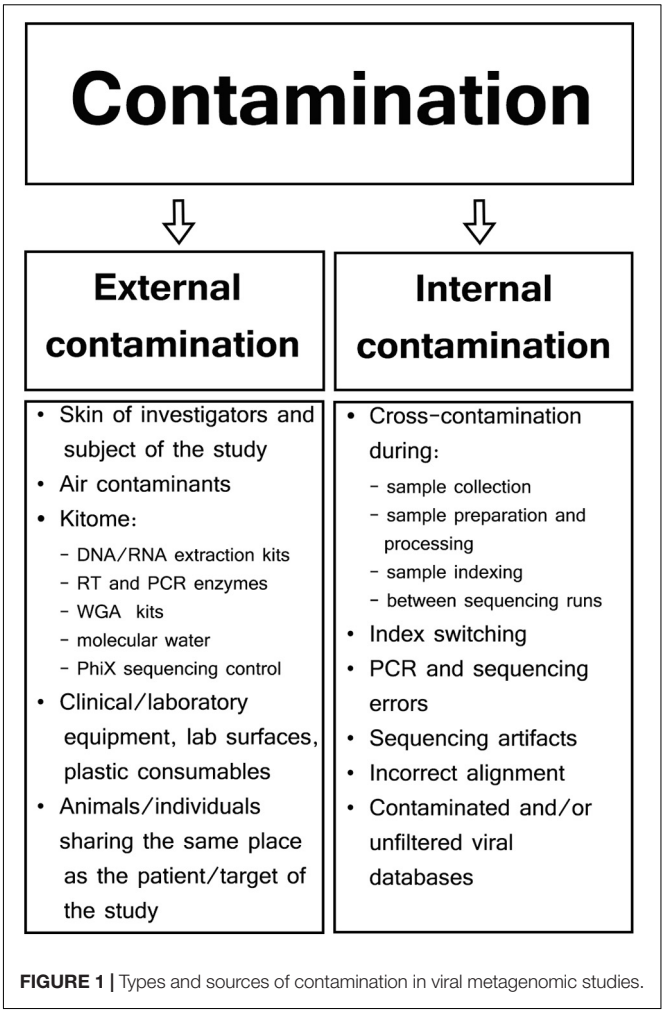

FIGURE 1 | Types and sources of contamination in viral metagenomic studies.

External Contamination

External contamination originates from the outside of samples during specimen collection and preparation and can include skin of patients or investigators (Kitchin et al., 1990; Meadow et al., 2015), clinical and laboratory equipment (Mukherjee et al., 2015; Llamas et al., 2017), collection tubes (Motley et al., 2014), contaminated laboratory surfaces or air (Bittinger et al., 2014), extraction kits, polymerase chain reaction (PCR) reagents (Grahn et al., 2003; Tilburg et al., 2010; Salter et al., 2014), or even molecular biology-grade water (Nogami et al., 1998; Kulakov et al., 2002; Keki et al., 2013). Manufacturers usually do not guarantee the absence of contaminating DNA in their products, and those reagents/kits that are sold as sterile may contain low-abundance external DNA (van der Horst et al., 2013). Generally, most external contaminations in microbiome studies have their own unique profile specific to particular reagents and kits; therefore, they are often referred to as kitome and are largely undistinguishable from microbiome signals derived from analyzed samples (van der Zee et al., 2002; Salter et al., 2014; Sabatier et al., 2020). Although a specific kitome can be detected and characterized, the types and quantities of reagent contaminants vary between different extraction/PCR kits and batches of the same reagent (Salter et al., 2014). True DNA/RNA signals are reproducible and associated with individual samples; however, reagent contamination signals are linked predominantly to specific batches or even reagents lots (Salter et al., 2014; de Goffau et al., 2018). For example, Glassing et al. (2016) analyzed MoBio DNA Extraction kit (QIAGEN; Hilden; Germany) and showed that 69% of dominating bacterial genera were the same in different lots of the kit, whereas the composition of minor genera was lot-dependent. Therefore, it has been recommended to process all samples in a particular project using the same batches/lots of reagents and to consider kit batches as a factor in the statistical analysis whenever multiple batches are used (Kim et al., 2017).

It seems that neither laboratories nor sequencing facilities are free from contamination, and this external DNA noise can change over time (Salter et al., 2014). For example, Weyrich et al. (2019) analyzed ultraclean ancient DNA laboratories for over 5 years and three modern molecular biology laboratories

for 1 year and found that each one had its own unique microbial profile that changed over time according to the month and season. In another study, three different laboratories performed 16S rRNA sequencing of the same *Salmonella bongori* control using different batches of the same extraction kits (FastDNA Spin Kit For Soil; MP Biomedicals, Santa Ana, CA, United States) and obtained three different microbial profiles. This variation in the contaminant content could be the result of differences between kit batches and other reagents or may represent contaminants specific for each laboratory environment and investigators (Salter et al., 2014; Kim et al., 2017).

Extraction kits seem to be the major source of nucleic acids external noise in microbiome studies (Evans et al., 2003; Salter et al., 2014; Smuts et al., 2014; Zhi et al., 2014; Sabatier et al., 2020). Glassing et al. (2016) identified 88 bacterial genera in commonly used DNA extraction kits, and it was estimated that 10–50% of the bacterial profiles in lower-airway human samples are contaminants, and their main source are extraction kits (Drengenes et al., 2019). Commercial extraction kits were found to contain a higher microbial diversity and several more human-associated bacterial taxa when compared to in-house extraction protocols (Weyrich et al., 2019). A different genetic background with significant higher prevalence of contaminants was reported for manual compared to automated extraction systems (Sabatier et al., 2020). The latter is not unexpected as manual extractions require a higher number of manual transfer steps than single-tube spin-column approach, and thus, the risk of external contamination is increased.

RNA sequencing is more susceptible to contamination than DNA sequencing due to the presence of the extra reverse transcription (RT) step (Strong et al., 2014). In addition, it was found that commercially available RT enzymes can contain viral contaminants such as equine infectious anemia virus or murine leukemia virus (MuLV); (Zheng et al., 2011; Wally et al., 2019).

DNA and RNA sequencing SM protocols may include an amplification step to generate sufficient amount of DNA/cDNA for sequencing libraries (Malboeuf et al., 2013). A number of studies documented the presence of external DNA in various commercial polymerases (Bottger, 1990; Schmidt et al., 1991; Hughes et al., 1994); for example, microbial contaminants were reported in six commercially available *Taq* polymerases (Iulia et al., 2013). It was estimated that the amounts of contaminants in recombinant *Taq* polymerase range between 10 and 1,000 genome equivalents of microbial DNA per unit of enzyme (Spangler et al., 2009). Other potential sources of contaminants could also include PCR buffers or $MgCl_2$ stocks, as well as primers prepared with water-containing contaminant DNA (Stinson et al., 2019). Considering the nature of SM and the necessity to analyze low-biomass samples, whole-genome amplification (WGA) is often used for the generation of templates suitable for sequencing (Thoendel et al., 2017). When three commercial WGA-DNA kits (Illustra V2 Genomiphi, Illustra single cell Genomiphi, and Qiagen REPLI-g single cell kits) were tested, it has been found that each contained a wide variety of microbial contaminant DNA (Thoendel et al., 2017). The origin of DNA background noise in WGA methods could come from amplification of contaminant DNA or from non-specific extension of random primers (Blainey and Quake, 2011). However, the consistent and highly specific contamination profile found in most individual WGA-DNA kits suggests the dominant role of the former (Thoendel et al., 2017). DNA background was reported in studies using WGA-RNA kits for the analysis of cerebrospinal fluid (CSF) and synovial fluid samples (Malboeuf et al., 2013; Perlejewski et al., 2015, 2016; Masters et al., 2018). WGA-RNA sequencing performed on clinical samples (CSF, swabs, and serum) and surrogate CSF samples (spiked with three 1:100 dilutions of influenza A H3N2 virus) using WTA2 kit (Sigma-Aldrich, St. Louis, MO, United States) resulted in the detection of a wide range of bacterial and viral contaminants. However, it should be noted that this background noise could have also originated from extraction kits and reagents used for the depletion of host genetic material (Oechslin et al., 2018).

The final step of wet-laboratory SM protocols is sequencing (Garmaeva et al., 2019). Currently, the most widely used method due to low costs, high yield, and wide availability is sequencing by synthesis marketed by Illumina (San Diego, CA, United States) (Kim et al., 2020). Despite numerous advantages, Illumina sequencing platforms share common challenge related to phage PhiX174 (approximately 5.3 kb) control used for quality and calibration assessment (Manley et al., 2016). While PhiX174 sequences should be removed from the final data, Mukherjee et al. (2015) reported that approximately 5.5% of publicly available microbial genomes in the Integrated Microbial Genomes database are contaminated by PhiX174, and 10% of them has been published in peer-reviewed scientific papers.

Internal Contamination

Cross-contamination is the most challenging internal contamination source when compared among the other numerous sources of internal contamination encountered in microbial sequencing (Olomu et al., 2020). This form of contamination results from transfer of genetic material, amplicons, or barcodes between reaction tubes or wells (Carlsen et al., 2012; Poore et al., 2020). Sample cross-contamination can occur at different steps throughout the whole SM protocol because of incorrect pipetting, accidental splashes of liquids, generation of aerosols, incorrect tube opening, or plate cover removal (Tamariz et al., 2006; Joung et al., 2017). The risk of sample cross-contamination increases when a large batch of samples undergoes extraction and/or library preparation, especially when using tube strips without individual caps, or when using reaction plates (Lejal et al., 2020; Olomu et al., 2020). Specimen-to-specimen cross-contamination was found to be significantly more common in high-throughput whole-genome sequencing (HT-WGS) in comparison to Sanger sequencing when influenza A/H3N2 virus from nasal/nasopharyngeal/throat swabs was analyzed (Lee et al., 2016). Well-to-well contamination affects primarily neighboring samples, but occasionally even those 10 wells apart (Minich et al., 2019). In a study conducted by Minich et al. (2019) on no-template controls (NTCs), 47.5% of blanks for tubes and 95.7% of blanks for plate DNA extractions had evidence of well-to-well contamination. This contaminating effect was more common in samples with low biomass, thus

negatively affecting microbial alpha and beta diversity metrics (Minich et al., 2019). To limit well-to-well contamination, it was proposed to keep a minimum of four-well gap between high- and low-biomass samples (Olomu et al., 2020).

Another type of cross-contamination is run-to-run contamination observed for MiSeq (Illumina, San Diego, CA, United States) sequencers, which may manifest itself for as many as seven sequential runs following the original run (Brumme and Poon, 2017; Eisenhofer et al., 2019). However, modifications to the post run wash procedure, mainly via the addition of a bleach wash, largely solved this problem (Brumme and Poon, 2017).

Another type of internal contamination occurs as a phenomenon called "index hopping" or "index switching" and is the main cause of incorrect sample assignment of sequencing reads in multiplexed pooled libraries (Griffiths et al., 2018). Index hopping refers to incorrect read assignment from a given NGS library based on assignment to a barcode belonging to a different one sequenced in the same pool (Costello et al., 2018). This effect is largely due to an excess of free index primers, which, together with the cluster generation reagents, randomly ligate to other samples pooled together in the sequencing run (Carlsen et al., 2012; Sinha et al., 2017; Costello et al., 2018). According to Sinha et al. (2017) in a multiplexed pool of samples sequenced on Illumina platform HiSeq 4000, up to 5–10% of all sequencing reads are misassigned from one sample to another. Index hopping is also a well-known phenomenon reported for the MinION (Oxford Nanopore Technologies, Oxford, Great Britain) sequencer where 0.056% of reads were found to have incorrectly assigned barcodes (Xu et al., 2018). Index switching reduces the value of negative controls in sequencing runs as NTCs and analyzed samples may contain the same sequences; thus, true signals cannot be distinguished from background noise (Hornung et al., 2019). To reduce index switching, unique dual-indexing and dual-matched indexed adapters with unique molecular indices are recommended (MacConaill et al., 2018; van der Valk et al., 2020).

Internal contamination in microbiome sequencing could also be caused by DNA damage and polymerase errors (Brandariz-Fontes et al., 2015; Hornung et al., 2019). In one study evaluating 13 commercial polymerases, it was found that enzyme choice has a large impact on the proportion of correct reads recovered from multiple gene sequencing run (from 17 to 71%) (Brandariz-Fontes et al., 2015). Nucleotide misincorporation, generation of chimeric sequences, or variation in efficiency of amplification of high and low GC fragments can arise from amplification bias (Brodin et al., 2013; Shugay et al., 2014). Sequencing of GC-poor regions on Illumina platforms is typically less efficient, which limits uniform read coverage across the genome, thus affecting viral genome assembly in SM analysis (Kozarewa et al., 2009; Chen Y. C. et al., 2013). A partial solution for amplification errors is offered by the use of high-fidelity polymerases, which are characterized by up to 100 times lower error rates and lower chimera generation rates (Sze and Schloss, 2019). Importantly, PCR conditions also play a significant role in generation of internal contaminants, and it has been demonstrated that a lower number of PCR cycles results in a lower signal-to-noise ratio in microbial profiling studies (Quail et al., 2011; Sze and Schloss, 2019). Errors can also arise during various parts of sequencing procedure (cluster amplification, sequencing cycles, image analysis), resulting in error base calling of approximately 0.1–1%, depending on sequencing platform (Fox et al., 2014).

Finally, internal contamination may be the result of erroneous bioinformatics reads classification (Hornung et al., 2019; Ye et al., 2019). In the current literature, there are dozens of virus-specific classification workflows that are based on different strategies using anything from simple one-step tools to analyses that combine five or more steps and a variety of algorithms for virome analysis (Wommack et al., 2012; Zhao et al., 2017; Nooij et al., 2018; Kieft et al., 2020). Nooij et al. (2018) evaluated 49 different workflows suitable for viral classification and found that the positive predictive value was generally high (>75%), although some classifiers had lower and varied precision scores: IMSA+A (9%), Kraken (34%), NBC (49%), and vFam (3–73%). Taxonomic classifiers are also associated with different default thresholds for false-positive signal detection (from 0.01 to 0.5%), which results in tens (Bracken, MetaPhlAn2) or even thousands (Centrifuge, CLARK, Kaiju, MMseqs2, PathSeq) of false-positive hits, depending on the workflow (Ye et al., 2019).

False reads assignments in microbiome studies may also arise from contamination of publicly available databases. For instance, more than 330,000 bp in the reference genomes of *Plasmodium gaboni* and *Plasmodium falciparum* were found to be contaminated with human genome (Kryukov and Imanishi, 2016). Merchant et al. (2014) discovered that *Neisseria gonorrhoeae* TCDC-NG08107 genome submitted to GenBank contained fragments of cattle and sheep genomes. Similarly, The Cancer Genome Atlas database was found to be contaminated by human papillomavirus type 38 originating from RNA sequencing of human endometrial samples (Kazemian et al., 2015). The previously mentioned contamination with PhiX174 illustrates the scale and range of microbial database contamination (Mukherjee et al., 2015).

CONTAMINATION IN LOW-BIOMASS SAMPLES

The impact of contamination is especially significant for low microbial biomass samples where the proportion of background noise increases with the decrease of input template (Malboeuf et al., 2013; Karstens et al., 2019). The quantity of biomass can be evaluated by comparing the amount of extracted DNA/RNA from the studied sample to the volume of genetic material isolated from NTCs in the same SM protocol. Samples specified as low biomass typically contain the amount of DNA/RNA similar to NTCs, whereas rich samples contain significantly more genetic material than blank controls (Lauder et al., 2016). Contaminants can easily dominate in low-biomass samples generating background noise that is much higher than true signal originating from the target virus (Malboeuf et al., 2013; Salter et al., 2014). So far, a wide variety of environmental and clinical samples containing low viral biomasses have been studied with SM workflows including air (Prussin et al., 2019), glacier ice (Zhong et al., 2020),

human skin (Tirosh et al., 2018), nasal swabs (Altan et al., 2019), and CSF (Perlejewski et al., 2020b; Perlejewski et al., 2020c). Most widely used library preparation kits for sequencing require inputs as low as 1 ng of DNA (e.g., llumina Nextra XT), but this may still be unattainable for some low-biomass samples. Based on our own experience and other published studies, some biological samples such as CSF yield <1 ng of DNA/RNA after typical 200 μL extraction, and viral load found in this material is often in the range of 100 copies/mL (Poissy et al., 2012; Bradshaw and Venkatesan, 2016; Perlejewski et al., 2020b). According to estimations by Garmaeva et al. (2019) 1 g of stool yields between 0.22 and 0.87 ng/μL of bacteriophage DNA (when using 50–200 μL of elution volume) signaling the need for application of more robust extraction and amplification methods, generating up to picograms of DNA (Garmaeva et al., 2019). To generate sufficient amount of material for library preparation and sequencing, new products based on techniques, such as multiple displacement amplification (Spits et al., 2006), linker amplification shotgun libraries (Bittinger et al., 2014), sequence-independent single-primer amplification (Chrzastek et al., 2017), or single-primer isothermal linear amplification (Ribo-SPIA), were introduced (Dafforn et al., 2004). Commercial kits such as NuGEN's Ovation RNA-Seq System V2, which is based on Ribo-SPIA, can generate sufficient amount of cDNA for library preparation from as little as 500 pg of RNA with sufficient coverage and read count when sequencing as little as 100 copies of HIV RNA (Malboeuf et al., 2013). Although all these methods solve the problem of insufficient material for sequencing in low-biomass samples, they neither reduce nor distinguish contaminants from true signals. Moreover, as previously mentioned, there is some evidence that these kits can be the source of extra genetic background themselves (Thoendel et al., 2017; Oechslin et al., 2018; Perlejewski et al., 2020a).

Another common problem specific for ultralow-biomass samples (input < 50 pg) is the high level of read duplication reaching 70%, whereas it was reported to be only 0.5–2% with high DNA input samples (>50 ng) (Garmaeva et al., 2019). This may generate a significant bias in quantitative analysis when different communities are compared to each other, and more irreproducible background noise is being amplified with decreasing sample biomass (Salter et al., 2014; Garmaeva et al., 2019; Erb-Downward et al., 2020). Finally, low-biomass samples require extra steps during extraction and library preparation, which increase the likelihood of external and internal contamination (Salter et al., 2014; Rawlinson et al., 2019).

CONTAMINANTS DETECTED IN VIRAL STUDIES

Viral Contaminants

Viral contaminants seem to be highly relevant among all types of contaminants encountered in viral metagenomic research, and they can, occasionally, significantly impact results interpretation, as was the case in the study by Xu et al. (2013) who identified National Institutes of Health–Chongqing virus (NIH-CQV) in patients with seronegative hepatitis. Although this particular pathogen was detected in 70% of hepatitis patients and in 0% of 45 healthy controls, it was later determined that this novel hybrid parvovirus-like virus was a contaminant from silica column–based RNA extraction kit (QIAamp MinElute Virus Kit; Qiagen, Hilden, Germany) (Smuts et al., 2014). The observed lack of NIH-CQV presence in healthy control subjects was probably related to lot-to-lot differences in the degree of spin column contamination (Naccache et al., 2014b). A year later, *Acanthocystis turfacea* chlorella virus 1 (ATCV-1) was proposed to be linked with the cognitive decline in humans after it was found in oropharyngeal samples collected from adults without current and past psychiatric disorders within a study that included measures of cognitive functioning (Yolken et al., 2014). ATCV-1 is of algal origin and was later found to be a part of kitome of commercial DNase and RNA extraction kits (Kjartansdottir et al., 2015). In general, most of reported contaminants in viral metagenomic studies seem to come from DNA and RNA extraction kits (Asplund et al., 2019).

In another study, a silica column–based kit (QIAamp viral RNA mini kit; Germany) was found to generate background noise of *Iridoviridae, Circoviridae, Baculoviridae,* and *Genomoviridae* sequences (Ngoi et al., 2016). In a recent study comparison of three extraction kits for metagenomic analysis of respiratory viruses, 19, 28, and 55 viral families were identified in NTCs using eMAG (bioMérieux, Marcy-l'Étoile, France), MagNA Pure 24 (Roche, Basel, Switzerland), and QIAamp Viral RNA Mini Kit extraction (Qiagen), respectively. Once again, the highest genetic background was found for the Qiagen kit, and it was composed of hits classified as *Siphoviridae, Myoviridae, Microviridae,* and *Podoviridae* (Sabatier et al., 2020). Various other reagents were also found to be a potential source of contamination, for example, BVDV−3 (bovine viral diarrhea virus 3) is a common contaminant in fetal bovine serum (Bergner et al., 2019), whereas MuLV is present in 17 human cell lines (Cao et al., 2015; Uphoff et al., 2015), as well as in reverse transcriptase–PCR reagents (Zheng et al., 2011; L'Huillier et al., 2019).

Separation of true signals from background noise is both extremely important and yet challenging in clinical settings. Bacteriophages are particularly common among a variety of contaminants encountered in clinical metagenomics (Naccache et al., 2014a; Ngoi et al., 2016; Asplund et al., 2019; Sangiovanni et al., 2019) and may disrupt the quantitative picture of virome, whereas sequences of eukaryotic viruses may be falsely associated with diseases (Xu et al., 2013). That was the case in the study linking febrile Kenyan adults with Kadiprio virus, which was initially considered to be the causative agent but was eventually found to be a part of QIAamp Viral RNA Mini Kit (Qiagen) kitome (Ngoi et al., 2016). In a recent study, Mollerup et al. (2019) used NGS to search for viruses in human cancers and found Merkel cell polyomavirus (MCPyV) in Merkel cell carcinomas. However, close similarity of all MCPyV sequences found across samples allowed studies to conclude laboratory surfaces as the source of contamination (Foulongne et al., 2011; Mollerup et al., 2019). In our previous viral SM studies, we often found pandoravirus sequences in CSF of patients with encephalitis and in NTCs (Perlejewski et al., 2015; Bukowska-Osko et al., 2016; Moustafa et al., 2017). After closer analysis of these sequences (low-complexity reads with nucleotide tandem repeats), they were determined not to represent true signals, but sequencing

artifacts and/or contaminants originating in laboratory reagents (Hjelmso et al., 2017; Waldvogel-Abramowski et al., 2019).

So far, there are very few studies addressing the issue of viral contamination in viral sequencing (Naccache et al., 2014b; Moustafa et al., 2017; Asplund et al., 2019). The most comprehensive was the one by Asplund et al. (2019) that evaluated 712 sequencing libraries prepared using several different protocols and found almost 500 viral hits associated with laboratory components. Similar to our observations, more contaminants were present in RNA sequencing protocols than those using DNA as a starting material. Most viruses reported by Asplund et al. (2019) were bacteriophages (60%), which is also consistent with our own studies in which phages constituted 96 and 77% of all viral sequences in CSF from encephalitis patients using RNA-based and DNA-based SM workflows, respectively (Perlejewski et al., 2020b). Viruses of non-human vertebrae hosts constituted approximately 12% of all viral contaminants (Asplund et al., 2019).

A frequent problem in viral SM studies is cross-contamination occurring when high viral-titer samples are simultaneously sequenced with low-biomass samples in the same sequencing lane (Moustafa et al., 2017). This is especially relevant when viral SM is performed using clinical samples, and overexpressed viral hits from one sample affect the viromes of other specimens. High-titer samples commonly contaminate low-biomass samples in the same sequencing run, and the rate of cross-contamination on Illumina platforms was reported to be approximately 0.05% (Deng et al., 2020). In nanopore sequencing, cross-contamination occurs when low- and high-titer samples are pooled; to remedy these problems, it was proposed to batch samples together according to viral loads (Lewandowski et al., 2019).

Bacterial Contaminants

Bacterial contaminants affect both viral SM and ATM studies in a similar manner because of the same external origin of bacterial sequences, which are usually present in the kitome (Salter et al., 2014). In approximately 72% of virome samples, bacterial DNA is considered to be the most abundant contaminant. Surprisingly, a significantly higher bacterial background noise is present in virus enriched than in non-enriched metagenomic samples (Zolfo et al., 2019). These findings indicate that many virus-like particles (VLP)–targeting SM workflows fail in efficient virus enrichment and experience large contamination problems.

The predominant bacterial genera found in negative controls in ATM and SM studies are *Propionibacterium*, *Flavobacterium*, *Streptococcus*, *Burkholderia*, *Methylobacterium*, *Curvibacter*, *Ralstonia*, *Escherichia*, *Acinetobacter*, and *Stenotrophomonas* (Lauder et al., 2016; Weyrich et al., 2019). Salter et al. (2014) reported the presence of *Proteobacteria*, *Actinobacteria*, *Firmicutes*, *Bacteroidetes*, *Deinococcus-Thermus*, and *Acidobacteria* in blank controls in PCR-based 16S rRNA gene and SM studies. In a study using HT-WGS in six different sequencing centers, *Bradyrhizobium* was reported to be the most common bacterial contaminant genus (Laurence et al., 2014). Moreno-Gallego et al. (2019) found that more than 1% of bacterial reads identified in a fecal virome represented contamination and they belonged largely to *Firmicutes* phylum. This is compatible with the findings of Zolfo et al., who analyzed bacterial contaminants using measurements of bacterial small subunit ribosomal RNA gene (SSU rRNA). In 37 virome studies (analyzed environmental and human samples), SSU rRNA median ranged from 0 to 14.3% (approximately 1.2% per data set); (Zolfo et al., 2019).

Host/Human Contaminants

In HT-WGS studies of such clinical samples as stool or CSF, host genomic reads are an integral part of whole metagenomes (Nakamura et al., 2009; Perlejewski et al., 2020c). Some investigators name all host reads as contaminants, as these sequences mask true signals and reduce assay sensitivity for pathogen detection (Malboeuf et al., 2013; Charre et al., 2020; Heravi et al., 2020). Moreover, overrepresentation of host sequences in large NGS data sets can extend the process of data analysis and require high and costly computational powers (Hasan et al., 2016).

The majority of human/host reads in WGS studies derived from the actual sample constitute a part of true genetic background; however, they reduce the sensitivity and sequencing coverage in microbial sequencing studies, especially for low-biomass samples (Chiu and Miller, 2019; Pereira-Marques et al., 2019). Clinical SM studies revealed that in such human-derived samples as nasopharyngeal aspirate, serum, and brain tissue, up to ∼95–99% of raw NGS reads derive from human DNA (Yang et al., 2011; Lipowski et al., 2017). Consequently, without a significant host genomic depletion, viral genome coverage is likely to be low even when high viral loads are present (Luk et al., 2015). In clinical settings, the minimum viral–host read ratio needed for viral identification is highly variable and species/sample/workflow-dependent. For instance, viral/human mRNA ratio of 0.0005% led to the discovery of MCPyV (Feng et al., 2008), whereas viral/human RNA ratio was 0.0135% when a new arenavirus causing febrile illness was first identified in patients who received solid organ transplants from a single donor (Palacios et al., 2008). In low-biomass clinical samples, human DNA/RNA overwhelms viral signals, but a variety of host depletion methods can partially remedy the problem by decreasing the background noise up to 3,100-fold with negligible loss of target virus (Oechslin et al., 2018). Unfortunately, with the reduction of host genomic contamination, an increase of non-host contaminants is common, especially when kitome-related signals are being amplified (Salter et al., 2014; Oechslin et al., 2018). Finally, some VLP purification methods such as CsCl density gradient ultracentrifugation efficiently remove host-derived DNA, but at the same time discriminate against particular viruses, thus affecting quantitative virome measurements (Kleiner et al., 2015).

Other Contaminants

Bacterial and host-derived sequences are rarely reported in SM viral studies because NGS reads are often not aligned to comprehensive databases that include non-viral genomes. In SM studies on human nasopharyngeal samples and CSF, reads were mapping to plant, parasitic, fungal genomes, and even synthetic

TABLE 1 | Recommendation for reducing contamination in viral metagenomic studies.

	Recommendations
General practices	• Use sterile laboratory equipment: tubes, tips with filter, decontaminated racks, and machines • Wear disposable protective coats, gloves, and face masks • Always decontaminate working area • Perform wet-laboratory work under laminar flow hood • Perform all steps in dedicated laboratory areas: create separate preamplification, amplification, and postamplification sites • Minimize the number of investigators in a project and record which samples were handled by a given technician
Sampling	• Avoid cross-contamination during sample preparation • Be aware that caging multiple laboratory animals in the same space may influence their microbial composition • Collect samples in sterile tubes • Avoid contamination derived from the skin or breath of the investigator • Use rich-biomass samples Maximize the sample volume for extraction when using low-biomass material
Reagents and wet-laboratory procedures	• Use the same types of reagents during the whole project Record all batches and lot numbers of all reagents used in a project • Minimize the number of steps in wet-laboratory workflow • Use dedicated extraction kits for low-biomass samples with low elution volumes • Keep in mind that silica column–based nucleic acid extraction kits are associated with numerous contaminants • Use highly purified enzymes and polymerases with high fidelity • Minimize the number of PCR cycles during amplification • Avoid using multichannel pipettes, sample plates, and strips without separate caps • If necessary make gaps in plates between samples • Use VLP enrichment workflows • Analyze the same biological samples in repeats
Sequencing	• Sequence all samples in a given project in the same sequencing center • Use unique dual barcoding • Sequence samples with similar viral titters in the same run • Minimize the number of PCR cycles during indexing
Controls	• Use blank and negative controls during sample preparation and extraction • Use non-template controls if amplification step is included • Use a variety of positive-control titrations to verify the accuracy of metagenomic workflow
Data analysis	• Create a list of contaminants specific for your viral metagenomic workflow and laboratory • Set your own threshold for contamination detection based on your results and experience • Align NGS reads to host and bacterial genomes to examine potential contamination • Set criteria for viral detection that include matching different regions of the viral genome with sufficient genome coverage • Align contigs rather than single NGS reads to viral genomes • Check the complexity of identified viral sequences to distinguish true signals from artifacts • Take into consideration sequencing error • Use verified and filtered viral databases for viral classification • Remove PhiX phage sequences before data upload • Use open-source decontamination software • Use dedicated software for viral detection and phage identification
Data interpretation and good practices	• For clinical diagnostic application, verify all potentially causative viral agents found in SM studies using PCRs • Pay close attention and be critical with regards to non-vertebrae viruses found in virome of vertebrae hosts • Perform batch/study/investigator associations with contaminants found in your data

constructs (Nakamura et al., 2009; Perlejewski et al., 2015). These hits could have derived from various sources including reagents, sequencing errors, and erroneous classification, especially when using unfiltered and biased genome databases for alignment.

CRITERIA FOR VIRUS IDENTIFICATION AND SEQUENCE DECONTAMINATION

In virus-targeted SM studies, it is critical to make an accurate distinction between true viral signals and contaminants (Xu et al., 2018; Asplund et al., 2019). This is especially difficult when low-biomass samples containing low viral loads are being analyzed (Malboeuf et al., 2013; Perlejewski et al., 2016). So far, a variety of SM workflows have been used for various samples using numerous wet-laboratory procedures and bioinformatics analysis, but a universally efficient approach is still unclear (Nakamura et al., 2009; Conceicao-Neto et al., 2015; Lewandowski et al., 2019).

SM viral protocols require validation and standardization before they can be used for routine clinical application

(van Boheemen et al., 2020). The protocols used are highly dependent on the type of sample. For instance, stool and tissue samples are treated differently (homogenization, filtration, DNA/RNA extraction, or nuclease treatment) than low-biomass samples such as CSF, human skin, or nasal swabs (e.g., required preamplification steps) (Hall et al., 2014; Sabatier et al., 2020). Thus, any future standardized SM clinical viral protocols must take into consideration sample type and the expected viral pathogen (either DNA or RNA-based approach) (Schlaberg et al., 2017; Kufner et al., 2019). Moreover, the same factors may affect the decision on sequencing parameters such as sequencing depth, which specifies how many times each base in a genome should be covered by NGS reads (Deng et al., 2020). This parameter is associated with the abundance of target virus, which affects the sensitivity of applied workflows (Malboeuf et al., 2013; Pereira-Marques et al., 2019). Another factor to consider is sequencing breadth, which specifies what portion of a genome should be sequenced for a reliable identification (Wylie et al., 2018). Ladner et al. (2014) proposed five categories to define different genome standards in viral-targeted sequencing beginning with a "standard draft," representing a low coverage with at least 50% of a draft genome candidate recovered (frequent for low-biomass samples with low viral loads). On the opposite site, a "finished" category requires high coverage rates (400–1,000×) and represents cases when a complete viral consensus genome sequence is obtained, combined with complete population-level characterization of genomic diversity (Ladner et al., 2014).

So far, there are no universal criteria for positive virus species identification in HTS-WGS analyses. Currently, it seems that the gold standard for microbial confirmation after identification by metagenomics is PCR or Sanger sequencing (Yu et al., 2016; Fang et al., 2018; Wylie et al., 2018; Holmes, 2019). Theoretically, even a one virus-specific NGS read in SM could indicate a true signal. In the already mentioned study, a novel arenavirus was identified in organ transplant setting after only 14 virus-specific sequences were detected by SM (Palacios et al., 2008). Liu et al. (2020) proposed that a positively identified viral taxon should be represented by at least two unique sequencing reads detected by the same or a different technique, whereas detection of reads mapping to at least three non-overlapping genome regions was required to identify virus in CSF in the studies conducted by Schlaberg et al. (2017) or Miller et al. (2019). Reads dispersed across the whole genome and with high coverage indicate the presence of true viral signals, but isolated and/or repeated viral sequences found across samples from the same run suggest sequencing artifacts (Asplund et al., 2019). In a study evaluating viral SM workflow in a tertiary diagnostic unit, positive viral identification required detection of at least three viral reads distributed across the whole genome with a high coverage score. Furthermore, the number of reads for the target virus had to be at least 100 times higher than in negative controls and other samples (Kufner et al., 2019). This approach is balanced as it takes into account the high possibility of cross-contamination between samples and NTCs, whereas many microbiome studies disqualified all sequences found in negative controls (Dunn et al., 2013; Karstens et al., 2019). A blacklist method assembles a catalog of specific contaminants found in NTCs in a given study and/or sequencing center and uses them in an algorithm to exclude matching sequences from WGS data sets (Ye et al., 2019). However, it is well-documented that true signals can also occur in NTCs as part of the index switching phenomenon (Callahan et al., 2017; Sinha et al., 2017; Costello et al., 2018; Larsson et al., 2018). It was shown that index switching ratios are higher in NTCs than in template-containing samples, indicating that at least several NTCs should be included in each sequencing run (Asplund et al., 2019). This approach allows for the detection of even sporadic contaminants, which is relevant if the decontamination is based on removal of sequences below a specified read/species abundance threshold (Lazarevic et al., 2016; Asplund et al., 2019).

Different thresholds were used in SM viral studies to distinguish between true and false-positive hits; for example, Guerin et al. (2020) proposed a threshold of >100 hits. In a study by Wylie et al. (2018) using pools of clinical samples (CSF, blood, plasma urine, swabs), the threshold of 0.1% of total reads for each virus expected in the appropriate sequencing pool was applied to limit the impact of index switching. In another study using VLP enrichment protocols, a relative read count threshold of 0.01% was set based on an empirical index contamination rate (O'Flaherty et al., 2018).

Viral identification is currently supported by numerous computational algorithms and open-source programs, such as VirSorter (Roux et al., 2015), VirusFinder (Wang et al., 2013), VirusSeeker (Zhao et al., 2017), VirusSeq (Chen Y. et al., 2013), VirusDetect (Zheng et al., 2017), and ViromeScan (Rampelli and Turroni, 2018). Some of the algorithms/pipelines [ViralFusionSeq (Li et al., 2013), Virana (Schelhorn et al., 2013), VERSE (Wang et al., 2015)] even allow for the detection of viruses integrated into the host genomes. Another group of useful programs such as MARVEL (Amgarten et al., 2018), PhagePhisher (Hatzopoulos et al., 2016), or Phage_Finder (Fouts, 2006) are designed to detect phages in metagenomic data sets. Special caution is required when interpreting the results of viral mining software applied in mixed metagenomes as they contain more computationally derived internal contamination compared to virus-specific data sets. Zolfo et al. (2019) showed that assembly carried out in poorly enriched metagenomes increases the number of contigs falsely classified as viral. More than 20% of assembled reads were assigned as viral in approximately 12% of metagenomic poorly enriched samples. This indicates a significant presence of viral false-positives found in data sets containing high representation of bacterial genomes (Zolfo et al., 2019).

Contamination in metagenomic studies can also be reduced or even removed using open-source software, such as R package decontam, which takes advantage of two observations: (i) contaminants are found at higher frequencies in low-titer samples, and (ii) their presence is more common in negative controls than in true samples (Davis et al., 2018). A similar application presents DecontaMiner, which uses a subtraction approach to detect contaminations by bacteria, fungi, and viruses from different sources (Sangiovanni et al., 2019). A much more virome-focused software is ViromeQC, which is designed for benchmarking and quantifying non-viral contamination in VLP-enriched projects. It uses three microbial markers: SSU-rRNA, large subunit rRNA gene, and 31 prokaryotic single-copy

markers. In addition, ViromeQC calculates viral enrichment score measuring the quality of VLP enrichment protocol (Zolfo et al., 2019). Finally, R packages such as microDecon (McKnight et al., 2019) or CroCo (Simion et al., 2018) are designed to efficiently and correctly detect cases of cross-contamination in studies using metabarcoding.

CONCLUDING REMARKS

Evolution of NGS and WGA methods has allowed for the development of numerous metagenomic workflows, which were successfully applied in viral-focused studies across various environments (Conceicao-Neto et al., 2015; Kohl et al., 2015; Perlejewski et al., 2020b). Regardless of the specific viral SM protocol, contamination cannot be totally avoided, and in particular, the issue of reagent contamination should always be addressed with high priority (Asplund et al., 2019). So far, the problem of contamination was mostly studied in 16S rRNA profiling, and only a few viral SM studies used NTCs or reported kitome sequences characteristic for their protocols (Grahn et al., 2003; Karstens et al., 2019).

In the present article, we described the most common sources and types of contamination found in viral metagenomic studies, and we propose some basic recommendations for reducing the background noise (**Table 1**). There is an urgent need for the development and validation of standards in viral metagenomics, which would limit contamination bias, increase the quality of research, and allow viral SM protocols to be more widely applied in diagnostics.

AUTHOR CONTRIBUTIONS

KP, HJ, and TP: writing—original draft preparation and visualization. KP and HJ: conceptualization, data curation, and writing—review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grant 2017/25/B/NZ6/01463 from the National Science Center, Poland (https://www.ncn.gov.pl/).

REFERENCES

Altan, E., Dib, J. C., Gulloso, A. R., Escribano Juandigua, D., Deng, X., Bruhn, R., et al. (2019). Effect of geographic isolation on the nasal virome of indigenous children. *J. Virol.* 93, e00681–19. doi: 10.1128/JVI.00681-19

Amgarten, D., Braga, L. P. P., Da Silva, A. M., and Setubal, J. C. (2018). MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Front. Genet.* 9:304. doi: 10.3389/fgene.2018.00304

Asplund, M., Kjartansdottir, K. R., Mollerup, 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

Bergner, L. M., Orton, R. J., Filipe, A. D., Shaw, A. E., Becker, D. J., Tello, C., et al. (2019). Using noninvasive metagenomics to characterize viral communities from wildlife. *Mol. Ecol. Resour.* 19, 128–143. doi: 10.1111/1755-0998.12946

Bittinger, K., Charlson, E. S., Loy, E., Shirley, D. J., Haas, A. R., Laughlin, A., et al. (2014). Improved characterization of medically relevant fungi in the human respiratory tract using next-generation sequencing. *Genome Biol.* 15:487. doi: 10.1186/s13059-014-0487-y

Blainey, P. C., and Quake, S. R. (2011). Digital MDA for enumeration of total nucleic acid contamination. *Nucleic Acids Res.* 39:e19. doi: 10.1093/nar/gkq1074

Boers, S. A., Jansen, R., and Hays, J. P. (2019). Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory. *Eur. J. Clin. Microbiol. Infect. Dis.* 38, 1059–1070. doi: 10.1007/s10096-019-03520-3

Bottger, E. C. (1990). Frequent contamination of Taq polymerase with DNA. *Clin. Chem.* 36, 1258–1259. doi: 10.1093/clinchem/36.6.1258b

Bradshaw, M. J., and Venkatesan, A. (2016). Herpes simplex virus-1 encephalitis in adults: pathophysiology, diagnosis, and management. *Neurotherapeutics* 13, 493–508. doi: 10.1007/s13311-016-0433-7

Brandariz-Fontes, C., Camacho-Sanchez, M., Vila, C., Vega-Pla, J. L., Rico, C., and Leonard, J. A. (2015). Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. *Sci. Rep.* 5:8056. doi: 10.1038/srep08056

Brodin, J., Mild, M., Hedskog, C., Sherwood, E., Leitner, T., Andersson, B., et al. (2013). PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. *PLoS One* 8:e70388. doi: 10.1371/journal.pone.0070388

Brumme, C. J., and Poon, A. F. Y. (2017). Promises and pitfalls of Illumina sequencing for HIV resistance genotyping. *Virus Res.* 239, 97–105. doi: 10.1016/j.virusres.2016.12.008

Bukowska-Osko, I., Perlejewski, K., Nakamura, S., Motooka, D., Stokowy, T., Kosinska, J., et al. (2016). Sensitivity of next-generation sequencing metagenomic analysis for detection of RNA and DNA viruses in cerebrospinal fluid: the confounding effect of background contamination. *Adv. Exp. Med. Biol.* 944, 53–62. doi: 10.1007/5584_2016_42

Callahan, B. J., Digiulio, D. B., Goltsman, D. S. A., Sun, C. L., Costello, E. K., Jeganathan, P., et al. (2017). Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women. *Proc. Natl. Acad. Sci. U.S.A.* 114, 9966–9971. doi: 10.1073/pnas.1705899114

Cao, S., Strong, M. J., Wang, X., Moss, W. N., Concha, M., Lin, Z., et al. (2015). High-throughput RNA sequencing-based virome analysis of 50 lymphoma cell lines from the cancer cell line encyclopedia project. *J. Virol.* 89, 713–729. doi: 10.1128/JVI.02570-14

Carlsen, T., Aas, A. B., Lindner, D., Vralstad, T., Schumacher, T., and Kauserud, H. (2012). Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecol.* 5, 747–749. doi: 10.1016/j.funeco.2012.06.003

Charre, C., Ginevra, C., Sabatier, M., Regue, H., Destras, G., Brun, S., et al. (2020). Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evol.* 6:veaa075. doi: 10.1093/ve/veaa075

Chen, Y., Yao, H., Thompson, E. J., Tannir, N. M., Weinstein, J. N., and Su, X. (2013). VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics* 29, 266–267. doi: 10.1093/bioinformatics/bts665

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., and Hwang, C. C. (2013). Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS One* 8:e62856. doi: 10.1371/journal.pone.0062856

Chiu, C. Y., and Miller, S. A. (2019). Clinical metagenomics. *Nat. Rev. Genet.* 20, 341–355. doi: 10.1038/s41576-019-0113-7

Chrzastek, K., Lee, D. H., Smith, D., Sharma, P., Suarez, D. L., Pantin-Jackwood, M., et al. (2017). Use of Sequence-Independent, Single-Primer-Amplification (SISPA) for rapid detection, identification, and characterization of avian RNA viruses. *Virology* 509, 159–166. doi: 10.1016/j.virol.2017.06.019

Conceicao-Neto, N., Zeller, M., Lefrere, H., De Bruyn, P., Beller, L., Deboutte, W., et al. (2015). Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci. Rep.* 5:16532. doi: 10.1038/srep16532

Costello, M., Fleharty, M., Abreu, J., Farjoun, Y., Ferriera, S., Holmes, L., et al. (2018). Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* 19:332. doi: 10.1186/s12864-018-4703-0

Dafforn, A., Chen, P., Deng, G., Herrler, M., Iglehart, D., Koritala, S., et al. (2004). Linear mRNA amplification from as little as 5 ng total RNA for global gene expression analysis. *Biotechniques* 37, 854–857. doi: 10.2144/04375PF01

Davis, N. M., Proctor, D. M., 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

de Goffau, M. C., Lager, S., Salter, S. J., Wagner, J., Kronbichler, A., Charnock-Jones, D. S., et al. (2018). Recognizing the reagent microbiome. *Nat. Microbiol.* 3, 851–853. doi: 10.1038/s41564-018-0202-y

Deng, X., Achari, A., Federman, S., Yu, G., Somasekar, S., Bartolo, I., et al. (2020). Author correction: metagenomic sequencing with spiked primer enrichment for viral diagnostics and genomic surveillance. *Nat. Microbiol.* 5:525. doi: 10.1038/s41564-020-0671-7

Drengenes, C., Wiker, H. G., Kalananthan, T., Nordeide, E., Eagan, T. M. L., and Nielsen, R. (2019). Laboratory contamination in airway microbiome studies. *BMC Microbiol.* 19:187. doi: 10.1186/s12866-019-1560-1

Dunn, R. R., Fierer, N., Henley, J. B., Leff, J. W., and Menninger, H. L. (2013). Home life: factors structuring the bacterial diversity found within and between homes. *PLoS One* 8:e64133. doi: 10.1371/journal.pone.0064133

Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R., and Weyrich, L. S. (2019). Contamination in low microbial biomass microbiome studies: issues and recommendations. *Trends Microbiol.* 27, 105–117. doi: 10.1016/j.tim.2018.11.003

Erb-Downward, J. R., Falkowski, N. R., D'souza, J. C., Mccloskey, L. M., Mcdonald, R. A., Brown, C. A., et al. (2020). Critical relevance of stochastic effects on low-bacterial-biomass 16S rRNA gene analysis. *mBio* 11, e00258–20. doi: 10.1128/mBio.00258-20

Evans, G. E., Murdoch, D. R., Anderson, T. P., Potter, H. C., George, P. M., and Chambers, S. T. (2003). Contamination of Qiagen DNA extraction kits with Legionella DNA. *J. Clin. Microbiol.* 41, 3452–3453. doi: 10.1128/JCM.41.7.3452-3453.2003

Fang, X., Xu, M., Fang, Q., Tan, H., Zhou, J., Li, Z., et al. (2018). Real-time utilization of metagenomic sequencing in the diagnosis and treatment monitoring of an invasive adenovirus B55 infection and subsequent herpes simplex virus encephalitis in an immunocompetent young adult. *Open Forum Infect. Dis.* 5:ofy114. doi: 10.1093/ofid/ofy114

Feng, H., Shuda, M., Chang, Y., and Moore, P. S. (2008). Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science* 319, 1096–1100. doi: 10.1126/science.1152586

Forbes, J. D., Knox, N. C., Ronholm, J., Pagotto, F., and Reimer, A. (2017). Metagenomics: the next culture-independent game changer. *Front. Microbiol.* 8:1069. doi: 10.3389/fmicb.2017.01069

Foulongne, V., Courgnaud, V., Champeau, W., and Segondy, M. (2011). Detection of Merkel cell polyomavirus on environmental surfaces. *J. Med. Virol.* 83, 1435–1439. doi: 10.1002/jmv.22110

Fouts, D. E. (2006). Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.* 34, 5839–5851. doi: 10.1093/nar/gkl732

Fox, E. J., Reid-Bayliss, K. S., Emond, M. J., and Loeb, L. A. (2014). Accuracy of next generation sequencing platforms. *Next Gener. Seq. Appl.* 1:1000106. doi: 10.4172/2469-9853.1000106

Garmaeva, S., Sinha, T., Kurilshikov, A., Fu, J., Wijmenga, C., and Zhernakova, A. (2019). Studying the gut virome in the metagenomic era: challenges and perspectives. *BMC Biol.* 17:84. doi: 10.1186/s12915-019-0704-y

Glassing, A., Dowd, S. E., Galandiuk, S., Davis, B., and Chiodini, R. J. (2016). Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog.* 8:24. doi: 10.1186/s13099-016-0103-7

Grahn, N., Olofsson, M., Ellnebo-Svedlund, K., Monstein, H. J., and Jonasson, J. (2003). Identification of mixed bacterial DNA contamination in broad-range PCR amplification of 16S rDNA V1 and V3 variable regions by pyrosequencing of cloned amplicons. *FEMS Microbiol. Lett.* 219, 87–91. doi: 10.1016/S0378-1097(02)01190-4

Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. L., and Marioni, J. C. (2018). Detection and removal of barcode swapping in single-cell RNA-seq data. *Nat. Commun.* 9:2667. doi: 10.1038/s41467-018-05083-x

Guerin, K., Rego, M., Bourges, D., Ersing, I., Haery, L., Harten Demaio, K., et al. (2020). A novel next-generation sequencing and analysis platform to assess the identity of recombinant adeno-associated viral preparations from viral DNA extracts. *Hum. Gene Ther.* 31, 664–678. doi: 10.1089/hum.2019.277

Hall, R. J., Wang, J., Todd, A. K., Bissielo, A. B., Yen, S., Strydom, H., et al. (2014). Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *J. Virol. Methods* 195, 194–204. doi: 10.1016/j.jviromet.2013.08.035

Hasan, M. R., Rawat, A., Tang, P., Jithesh, P. V., Thomas, E., Tan, R., et al. (2016). Depletion of human DNA in spiked clinical specimens for improvement of sensitivity of pathogen detection by next-generation sequencing. *J. Clin. Microbiol.* 54, 919–927. doi: 10.1128/JCM.03050-15

Hatzopoulos, T., Watkins, S. C., and Putonti, C. (2016). PhagePhisher: a pipeline for the discovery of covert viral sequences in complex genomic datasets. *Microb. Genom.* 2, e000053. doi: 10.1099/mgen.0.000053

Heravi, F. S., Zakrzewski, M., Vickery, K., and Hu, H. (2020). Host DNA depletion efficiency of microbiome DNA enrichment methods in infected tissue samples. *J. Microbiol. Methods* 170:105856. doi: 10.1016/j.mimet.2020.105856

Hjelmso, M. H., Hellmer, M., Fernandez-Cassi, X., Timoneda, N., Lukjancenko, O., Seidel, M., et al. (2017). Evaluation of methods for the concentration and extraction of viruses from sewage in the context of metagenomic sequencing. *PLoS One* 12:e0170199. doi: 10.1371/journal.pone.0170199

Holmes, E. C. (2019). Reagent contamination in viromics: all that glitters is not gold. *Clin. Microbiol. Infect.* 25, 1167–1168. doi: 10.1016/j.cmi.2019.06.019

Hornung, B. V. H., Zwittink, R. D., and Kuijper, E. J. (2019). Issues and current standards of controls in microbiome research. *FEMS Microbiol. Ecol.* 95:fiz045. doi: 10.1093/femsec/fiz045

Hughes, M. S., Beck, L. A., and Skuce, R. A. (1994). Identification and elimination of DNA sequences in Taq DNA polymerase. *J. Clin. Microbiol.* 32, 2007–2008. doi: 10.1128/jcm.32.8.2007-2008.1994

Huson, D. H., and Mitra, S. (2012). Introduction to the analysis of environmental sequences: metagenomics with MEGAN. *Methods Mol. Biol.* 856, 415–429. doi: 10.1007/978-1-61779-585-5_17

Iulia, L., Bianca, I. M., Cornelia, O., and Octavian, P. (2013). The evidence of contaminant bacterial DNA in several commercial Taq polymerases. *Rom. Biotechnol. Lett.* 18, 8007–8012.

Joung, Y. S., Ge, Z., and Buie, C. R. (2017). Bioaerosol generation by raindrops on soil. *Nat. Commun.* 8:14668. doi: 10.1038/ncomms14668

Karstens, L., Asquith, M., Davin, S., Fair, D., Gregory, W. T., Wolfe, A. J., et al. (2019). Controlling for contaminants in low-biomass 16S rRNA gene sequencing experiments. *mSystems* 4, e00290–19. doi: 10.1128/mSystems.00290-19

Kazemian, M., Ren, M., Lin, J. X., Liao, W., Spolski, R., and Leonard, W. J. (2015). Possible human papillomavirus 38 contamination of endometrial cancer RNA sequencing samples in the cancer genome atlas database. *J. Virol.* 89, 8967–8973. doi: 10.1128/JVI.00822-15

Keki, Z., Grebner, K., Bohus, V., Marialigeti, K., and Toth, E. M. (2013). Application of special oligotrophic media for cultivation of bacterial communities originated from ultrapure water. *Acta Microbiol. Immunol. Hung.* 60, 345–357. doi: 10.1556/AMicr.60.2013.3.9

Kieft, K., Zhou, Z., and Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 8:90. doi: 10.1186/s40168-020-00867-0

Kim, D., Hofstaedter, C. E., Zhao, C., Mattei, L., Tanes, C., Clarke, E., et al. (2017). Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* 5:52. doi: 10.1186/s40168-017-0267-5

Kim, S., Lee, J. W., and Park, Y. S. (2020). The application of next-generation sequencing to define factors related to oral cancer and discover novel biomarkers. *Life (Basel)* 10:228. doi: 10.3390/life10100228

Kitchin, P. A., Szotyori, Z., Fromholc, C., and Almond, N. (1990). Avoidance of PCR false positives [corrected]. *Nature* 344:201. doi: 10.1038/344201a0

Kjartansdottir, K. R., Friis-Nielsen, J., Asplund, M., Mollerup, S., Mourier, T., Jensen, R. H., et al. (2015). Traces of ATCV-1 associated with laboratory

component contamination. *Proc. Natl. Acad. Sci. U.S.A.* 112, E925–E926. doi: 10.1073/pnas.1423756112

Kleiner, M., Hooper, L. V., and Duerkop, B. A. (2015). Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics* 16:7. doi: 10.1186/s12864-014-1207-4

Kohl, C., Brinkmann, A., Dabrowski, P. W., Radonic, A., Nitsche, A., and Kurth, A. (2015). Protocol for metagenomic virus detection in clinical specimens. *Emerg. Infect. Dis.* 21, 48–57. doi: 10.3201/eid2101.140766

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., and Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* 6, 291–295. doi: 10.1038/nmeth.1311

Kryukov, K., and Imanishi, T. (2016). Human contamination in public genome assemblies. *PLoS One* 11:e0162424. doi: 10.1371/journal.pone.0162424

Kufner, V., Plate, A., Schmutz, S., Braun, D. L., Gunthard, H. F., Capaul, R., et al. (2019). Two years of viral metagenomics in a tertiary diagnostics unit: evaluation of the first 105 cases. *Genes (Basel)* 10:661. doi: 10.3390/genes10090661

Kulakov, L. A., Mcalister, M. B., Ogden, K. L., Larkin, M. J., and O'hanlon, J. F. (2002). Analysis of bacteria contaminating ultrapure water in industrial systems. *Appl. Environ. Microbiol.* 68, 1548–1555. doi: 10.1128/AEM.68.4.1548-1555.2002

Ladner, J. T., Beitzel, B., Chain, P. S., Davenport, M. G., Donaldson, E. F., Frieman, M., et al. (2014). Standards for sequencing viral genomes in the era of high-throughput sequencing. *mBio* 5, e01360–14. doi: 10.1128/mBio.01360-14

Larsson, A. J. M., Stanley, G., Sinha, R., Weissman, I. L., and Sandberg, R. (2018). Computational correction of index switching in multiplexed sequencing libraries. *Nat. Methods* 15, 305–307. doi: 10.1038/nmeth.4666

Lauder, A. P., Roche, A. M., Sherrill-Mix, S., Bailey, A., Laughlin, A. L., Bittinger, K., et al. (2016). Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome* 4:29. doi: 10.1186/s40168-016-0172-3

Laurence, M., Hatzis, C., and Brash, D. E. (2014). Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS One* 9:e97876. doi: 10.1371/journal.pone.0097876

Lazarevic, V., Gaia, N., Girard, M., and Schrenzel, J. (2016). Decontamination of 16S rRNA gene amplicon sequence datasets based on bacterial load assessment by qPCR. *BMC Microbiol.* 16:73. doi: 10.1186/s12866-016-0689-4

Lee, H. K., Lee, C. K., Tang, J. W., Loh, T. P., and Koay, E. S. (2016). Contamination-controlled high-throughput whole genome sequencing for influenza A viruses using the MiSeq sequencer. *Sci. Rep.* 6:33318. doi: 10.1038/srep33318

Lejal, E., Estrada-Pena, A., Marsot, M., Cosson, J. F., Rue, O., Mariadassou, M., et al. (2020). Taxon appearance from extraction and amplification steps demonstrates the value of multiple controls in tick microbiota analysis. *Front. Microbiol.* 11:1093. doi: 10.3389/fmicb.2020.01093

Lewandowski, K., Xu, Y., Pullan, S. T., Lumley, S. F., Foster, D., Sanderson, N., et al. (2019). Metagenomic nanopore sequencing of influenza virus direct from clinical respiratory samples. *J. Clin. Microbiol.* 58, e00963–19. doi: 10.1128/JCM.00963-19

L'Huillier, A. G., Brito, F., Wagner, N., Cordey, S., Zdobnov, E., Posfay-Barbe, K. M., et al. (2019). Identification of viral signatures using high-throughput sequencing on blood of patients with Kawasaki disease. *Front. Pediatr.* 7:524. doi: 10.3389/fped.2019.00524

Li, J. W., Wan, R., Yu, C. S., Co, N. N., Wong, N., and Chan, T. F. (2013). ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. *Bioinformatics* 29, 649–651. doi: 10.1093/bioinformatics/btt011

Lipowski, D., Popiel, M., Perlejewski, K., Nakamura, S., Bukowska-Osko, I., Rzadkiewicz, E., et al. (2017). A cluster of fatal tick-borne encephalitis virus infection in organ transplant setting. *J. Infect. Dis.* 215, 896–901. doi: 10.1093/infdis/jix040

Liu, B., Shao, N., Wang, J., Zhou, S., Su, H., Dong, J., et al. (2020). An optimized metagenomic approach for virome detection of clinical pharyngeal samples with respiratory infection. *Front. Microbiol.* 11:1552. doi: 10.3389/fmicb.2020.01552

Llamas, B., Valverde, G., Fehren-Schmitz, L., Weyrich, L. S., Cooper, A., and Haak, W. (2017). From the field to the laboratory: controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. *Sci. Technol. Archaeol. Res.* 3, 1–14. doi: 10.1080/20548923.2016.1258824

Luk, K. C., Berg, M. G., Naccache, S. N., Kabre, B., Federman, S., Mbanya, D., et al. (2015). Utility of metagenomic next-generation sequencing for characterization of HIV and human pegivirus diversity. *PLoS One* 10:e0141723. doi: 10.1371/journal.pone.0141723

MacConaill, L. E., Burns, R. T., Nag, A., Coleman, H. A., Slevin, M. K., Giorda, K., et al. (2018). Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics* 19:30. doi: 10.1186/s12864-017-4428-5

Malboeuf, C. M., Yang, X., Charlebois, P., Qu, J., Berlin, A. M., Casali, M., et al. (2013). Complete viral RNA genome sequencing of ultra-low copy samples by sequence-independent amplification. *Nucleic Acids Res.* 41:e13. doi: 10.1093/nar/gks794

Manley, L. J., Ma, D., and Levine, S. S. (2016). Monitoring error rates in illumina sequencing. *J. Biomol. Tech.* 27, 125–128. doi: 10.7171/jbt.16-2704-002

Masters, T. L., Hilker, C. A., Jeraldo, P. R., Bhagwate, A. V., Greenwood-Quaintance, K. E., Eckloff, B. W., et al. (2018). Comparative evaluation of cDNA library construction approaches for RNA-Seq analysis from low RNA-content human specimens. *J. Microbiol. Methods* 154, 55–62. doi: 10.1016/j.mimet.2018.10.008

McKnight, D. T., Huerlimann, R., Bower, D. S., Schwarzkopf, L., Alford, R. A., and Zenger, K. R. (2019). microDecon: a highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies. *Environ. DNA* 1, 14–25. doi: 10.1002/edn3.11

Meadow, J. F., Altrichter, A. E., Bateman, A. C., Stenson, J., Brown, G. Z., Green, J. L., et al. (2015). Humans differ in their personal microbial cloud. *PeerJ* 3:e1258. doi: 10.7717/peerj.1258

Merchant, S., Wood, D. E., and Salzberg, S. L. (2014). Unexpected cross-species contamination in genome sequencing projects. *PeerJ* 2:e675. doi: 10.7717/peerj.675

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

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, 1–13. doi: 10.1128/mSystems.00186-19

Mollerup, S., Asplund, M., Friis-Nielsen, J., Kjartansdottir, K. R., Fridholm, H., Hansen, T. A., et al. (2019). High-throughput sequencing-based investigation of viruses in human cancers by multienrichment approach. *J. Infect. Dis.* 220, 1312–1324. doi: 10.1093/infdis/jiz318

Moreno-Gallego, J. L., Chou, S. P., Di Rienzi, S. C., Goodrich, J. K., Spector, T. D., Bell, J. T., et al. (2019). Virome diversity correlates with intestinal microbiome diversity in adult monozygotic twins. *Cell Host Microbe* 25, 261–272.e5. doi: 10.1016/j.chom.2019.01.019

Motley, S. T., Picuri, J. M., Crowder, C. D., Minich, J. J., Hofstadler, S. A., and Eshoo, M. W. (2014). Improved multiple displacement amplification (iMDA) and ultraclean reagents. *BMC Genomics* 15:443. doi: 10.1186/1471-2164-15-443

Moustafa, A., Xie, C., Kirkness, E., Biggs, W., Wong, E., Turpaz, Y., et al. (2017). The blood DNA virome in 8,000 humans. *PLoS Pathog.* 13:e1006292. doi: 10.1371/journal.ppat.1006292

Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C., and Pati, A. (2015). Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand. Genomic Sci.* 10:18. doi: 10.1186/1944-3277-10-18

Naccache, S. N., Hackett, J. Jr., Delwart, E. L., and Chiu, C. Y. (2014b). Concerns over the origin of NIH-CQV, a novel virus discovered in Chinese patients with seronegative hepatitis. *Proc. Natl. Acad. Sci. U.S.A.* 111:E976. doi: 10.1073/pnas.1317064111

Naccache, S. N., Federman, S., Veeraraghavan, N., Zaharia, M., Lee, D., Samayoa, E., et al. (2014a). A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res.* 24, 1180–1192. doi: 10.1101/gr.171934.113

Nakamura, S., Yang, C. S., Sakon, N., Ueda, M., Tougan, T., Yamashita, A., et al. (2009). Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS One* 4:e4219. doi: 10.1371/journal.pone.0004219

Ngoi, C. N., Siqueira, J., Li, L. L., Deng, X. T., Mugo, P., Graham, S. M., et al. (2016). The plasma virome of febrile adult Kenyans shows frequent parvovirus B19 infections and a novel arbovirus (Kadipiro virus). *J. Gen. Virol.* 97, 3359–3367. doi: 10.1099/jgv.0.000644

Nogami, T., Ohto, T., Kawaguchi, O., Zaitsu, Y., and Sasaki, S. (1998). Estimation of bacterial contamination in ultrapure water: application of the anti-DNA antibody. *Anal. Chem.* 70, 5296–5301. doi: 10.1021/ac9805854

Nooij, S., Schmitz, D., Vennema, H., Kroneman, A., and Koopmans, M. P. G. (2018). Overview of virus metagenomic classification methods and their biological applications. *Front. Microbiol.* 9:749. doi: 10.3389/fmicb.2018.00749

Oechslin, C. P., Lenz, N., Liechti, N., Ryter, S., Agyeman, P., Bruggmann, R., et al. (2018). Limited correlation of shotgun metagenomics following host depletion and routine diagnostics for viruses and bacteria in low concentrated surrogate and clinical samples. *Front. Cell. Infect. Microbiol.* 8:375. doi: 10.3389/fcimb.2018.00375

O'Flaherty, B. M., Li, Y., Tao, Y., Paden, C. R., Queen, K., Zhang, J., et al. (2018). Comprehensive viral enrichment enables sensitive respiratory virus genomic identification and analysis by next generation sequencing. *Genome Res.* 28, 869–877. doi: 10.1101/gr.226316.117

Olomu, I. N., Pena-Cortes, L. C., Long, R. A., Vyas, A., Krichevskiy, O., Luellwitz, R., 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

Palacios, G., Druce, J., Du, L., Tran, T., Birch, C., Briese, T., et al. (2008). A new arenavirus in a cluster of fatal transplant-associated diseases. *N. Engl. J. Med.* 358, 991–998. doi: 10.1056/NEJMoa073785

Pereira-Marques, J., Hout, A., Ferreira, R. M., Weber, M., Pinto-Ribeiro, I., Van Doorn, L. J., et al. (2019). Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Front. Microbiol.* 10:1277. doi: 10.3389/fmicb.2019.01277

Perlejewski, K., Bukowska-Osko, I., Nakamura, S., Motooka, D., Stokowy, T., Ploski, R., et al. (2016). Metagenomic analysis of cerebrospinal fluid from patients with multiple sclerosis. *Adv. Exp. Med. Biol.* 935, 89–98. doi: 10.1007/5584_2016_25

Perlejewski, K., Bukowska-Osko, I., Rydzanicz, M., Pawelczyk, A., Caraballo Corts, K., Osuch, S., et al. (2020b). Next-generation sequencing in the diagnosis of viral encephalitis: sensitivity and clinical limitations. *Sci. Rep.* 10:16173. doi: 10.1038/s41598-020-73156-3

Perlejewski, K., Bukowska-Osko, I., Rydzanicz, M., Dzieciatkowski, T., Zakrzewska-Pniewska, B., Podlecka-Pietowska, A., et al. (2020a). Search for viral agents in cerebrospinal fluid in patients with multiple sclerosis using real-time PCR and metagenomics. *PLoS One* 15:e0240601. doi: 10.1371/journal.pone.0240601

Perlejewski, K., Pawelczyk, A., Bukowska-Osko, I., Rydzanicz, M., Dzieciatkowski, T., Paciorek, M., et al. (2020c). Search for viral infections in cerebrospinal fluid from patients with autoimmune encephalitis. *Open Forum Infect. Dis.* 7:ofaa468. doi: 10.1093/ofid/ofaa468

Perlejewski, K., Popiel, M., Laskus, T., Nakamura, S., Motooka, D., Stokowy, T., et al. (2015). Next-Generation Sequencing (NGS) in the identification of encephalitis-causing viruses: unexpected detection of human herpesvirus 1 while searching for RNA pathogens. *J. Virol. Methods* 226, 1–6. doi: 10.1016/j.jviromet.2015.09.010

Poissy, J., Champenois, K., Dewilde, A., Melliez, H., Georges, H., Senneville, E., et al. (2012). Impact of Herpes simplex virus load and red blood cells in cerebrospinal fluid upon herpes simplex meningo-encephalitis outcome. *BMC Infect. Dis.* 12:356. doi: 10.1186/1471-2334-12-356

Poore, G. D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., et al. (2020). Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* 579, 567–574. doi: 10.1038/s41586-020-2095-1

Prussin, A. J. II, Torres, P. J., Shimashita, J., Head, S. R., Bibby, K. J., Kelley, S. T., et al. (2019). Seasonal dynamics of DNA and RNA viral bioaerosol communities in a daycare center. *Microbiome* 7:53. doi: 10.1186/s40168-019-0672-z

Quail, M. A., Otto, T. D., Gu, Y., Harris, S. R., Skelly, T. F., Mcquillan, J. A., et al. (2011). Optimal enzymes for amplifying sequencing libraries. *Nat. Methods* 9, 10–11. doi: 10.1038/nmeth.1814

Rampelli, S., and Turroni, S. (2018). From whole-genome shotgun sequencing to viral community profiling: the viromescan tool. *Methods Mol. Biol.* 1746, 181–185. doi: 10.1007/978-1-4939-7683-6_14

Rawlinson, S., Ciric, L., and Cloutman-Green, E. (2019). How to carry out microbiological sampling of healthcare environment surfaces? A review of current evidence. *J. Hosp. Infect.* 103, 363–374. doi: 10.1016/j.jhin.2019.07.015

Roux, S., Enault, F., Hurwitz, B. L., and Sullivan, M. B. (2015). VirSorter: mining viral signal from microbial genomic data. *PeerJ* 3:e985. doi: 10.7717/peerj.985

Sabatier, M., Bal, A., Destras, G., Regue, H., Queromes, G., Cheynet, V., et al. (2020). Comparison of nucleic acid extraction methods for a viral metagenomics analysis of respiratory viruses. *Microorganisms* 8:1539. doi: 10.3390/microorganisms8101539

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

Sangiovanni, M., Granata, I., Thind, A. S., and Guarracino, M. R. (2019). From trash to treasure: detecting unexpected contamination in unmapped NGS data. *BMC Bioinformatics* 20:168. doi: 10.1186/s12859-019-2684-x

Schelhorn, S. E., Fischer, M., Tolosi, L., Altmuller, J., Nurnberg, P., Pfister, H., et al. (2013). Sensitive detection of viral transcripts in human tumor transcriptomes. *PLoS Comput. Biol.* 9:e1003228. doi: 10.1371/journal.pcbi.1003228

Schlaberg, R., Chiu, C. Y., Miller, S., Procop, G. W., Weinstock, G., and Professional Practice Committee and Committee on Laboratory Practices of the American Society for Microbiology (2017). Validation of metagenomic next-generation sequencing tests for universal pathogen detection. *Arch. Pathol. Lab. Med.* 141, 776–786. doi: 10.5858/arpa.2016-0539-RA

Schmidt, T. M., Pace, B., and Pace, N. R. (1991). Detection of DNA contamination in Taq polymerase. *Biotechniques* 11, 176–177.

Shugay, M., Britanova, O. V., Merzlyak, E. M., Turchaninova, M. A., Mamedov, I. Z., Tuganbaev, T. R., et al. (2014). Towards error-free profiling of immune repertoires. *Nat. Methods* 11, 653–655. doi: 10.1038/nmeth.2960

Simion, P., Belkhir, K., Francois, C., Veyssier, J., Rink, J. C., Manuel, M., et al. (2018). A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. *BMC Biol.* 16:28. doi: 10.1186/s12915-018-0486-7

Sinha, R., Stanley, G., Gulati, G. S., Ezran, C., Travaglini, K. J., Wei, E., et al. (2017). Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. *bioRxiv*[Preprint]. doi: 10.1101/125724

Smuts, H., Kew, M., Khan, A., and Korsman, S. (2014). Novel hybrid parvovirus-like virus, NIH-CQV/PHV, contaminants in silica column-based nucleic acid extraction kits. *J. Virol.* 88:1398. doi: 10.1128/JVI.03206-13

Spangler, R., Goddard, N. L., and Thaler, D. S. (2009). Optimizing Taq polymerase concentration for improved signal-to-noise in the broad range detection of low abundance bacteria. *PLoS One* 4:e7010. doi: 10.1371/journal.pone.0007010

Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I., et al. (2006). Whole-genome multiple displacement amplification from single cells. *Nat. Protoc.* 1, 1965–1970. doi: 10.1038/nprot.2006.326

Stinson, L. F., Keelan, J. A., and Payne, M. S. (2019). Identification and removal of contaminating microbial DNA from PCR reagents: impact on low-biomass microbiome analyses. *Lett. Appl. Microbiol.* 68, 2–8. doi: 10.1111/lam.13091

Strong, M. J., Xu, G., Morici, L., Splinter Bon-Durant, S., Baddoo, M., Lin, Z., et al. (2014). Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. *PLoS Pathog.* 10:e1004437. doi: 10.1371/journal.ppat.1004437

Sze, M. A., and Schloss, P. D. (2019). The impact of DNA polymerase and number of rounds of amplification in PCR on 16S rRNA gene sequence data. *mSphere* 4, 1–13. doi: 10.1128/mSphere.00163-19

Tamariz, J., Voynarovska, K., Prinz, M., and Caragine, T. (2006). The application of ultraviolet irradiation to exogenous sources of DNA in plasticware and water for the amplification of low copy number DNA. *J. Forensic Sci.* 51, 790–794. doi: 10.1111/j.1556-4029.2006.00172.x

Thoendel, M., Jeraldo, P., Greenwood-Quaintance, K. E., Yao, J., Chia, N., Hanssen, A. D., et al. (2017). Impact of contaminating DNA in whole-genome amplification kits used for metagenomic shotgun sequencing for infection diagnosis. *J. Clin. Microbiol.* 55, 1789–1801. doi: 10.1128/JCM.02402-16

Tilburg, J. J., Nabuurs-Franssen, M. H., Van Hannen, E. J., Horrevorts, A. M., Melchers, W. J., and Klaassen, C. H. (2010). Contamination of commercial PCR master mix with DNA from *Coxiella burnetii*. *J. Clin. Microbiol.* 48, 4634–4635. doi: 10.1128/JCM.00464-10

Tirosh, O., Conlan, S., Deming, C., Lee-Lin, S. Q., Huang, X., Program, N. C. S., et al. (2018). Expanded skin virome in DOCK8-deficient patients. *Nat. Med.* 24, 1815–1821. doi: 10.1038/s41591-018-0211-7

Uphoff, C. C., Lange, S., Denkmann, S. A., Garritsen, H. S., and Drexler, H. G. (2015). Prevalence and characterization of murine leukemia virus contamination in human cell lines. *PLoS One* 10:e0125622. doi: 10.1371/journal.pone.0125622

van Boheemen, S., Van Rijn, A. L., Pappas, N., Carbo, E. C., Vorderman, R. H. P., Sidorov, I., et al. (2020). Retrospective validation of a metagenomic sequencing protocol for combined detection of RNA and DNA viruses using respiratory samples from pediatric patients. *J. Mol. Diagn.* 22, 196–207. doi: 10.1016/j.jmoldx.2019.10.007

van der Horst, J., Buijs, M. J., Laine, M. L., Wismeijer, D., Loos, B. G., Crielaard, W., et al. (2013). Sterile paper points as a bacterial DNA-contamination source in microbiome profiles of clinical samples. *J. Dent.* 41, 1297–1301. doi: 10.1016/j.jdent.2013.10.008

van der Valk, T., Vezzi, F., Ormestad, M., Dalen, L., and Guschanski, K. (2020). Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. *Mol. Ecol. Resour.* 20, 1171–1181. doi: 10.1111/1755-0998.13009

van der Zee, A., Peeters, M., De Jong, C., Verbakel, H., Crielaard, J. W., Claas, E. C., et al. (2002). Qiagen DNA extraction kits for sample preparation for legionella PCR are not suitable for diagnostic purposes. *J. Clin. Microbiol.* 40:1126. doi: 10.1128/JCM.40.3.1128.2002

Waldvogel-Abramowski, S., Taleb, S., Alessandrini, M., and Preynat-Seauve, O. (2019). Viral metagenomics of blood donors and blood-derived products using next-generation sequencing. *Transfus. Med. Hemother.* 46, 87–93. doi: 10.1159/000499088

Wally, N., Schneider, M., Thannesberger, J., Kastner, M. T., Bakonyi, T., Indik, S., et al. (2019). Plasmid DNA contaminant in molecular reagents. *Sci. Rep.* 9:1652. doi: 10.1038/s41598-019-38733-1

Wang, Q., Jia, P., and Zhao, Z. (2013). VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PLoS One* 8:e64465. doi: 10.1371/journal.pone.0064465

Wang, Q., Jia, P., and Zhao, Z. (2015). VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. *Genome Med.* 7:2. doi: 10.1186/s13073-015-0126-6

Weyrich, L. S., Farrer, A. G., Eisenhofer, R., Arriola, L. A., Young, J., Selway, C. A., et al. (2019). Laboratory contamination over time during low-biomass sample analysis. *Mol. Ecol. Resour.* 19, 982–996. doi: 10.1111/1755-0998.13011

Wommack, K. E., Bhavsar, J., Polson, S. W., Chen, J., Dumas, M., Srinivasiah, S., et al. (2012). VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand. Genomic Sci.* 6, 427–439. doi: 10.4056/sigs.2945050

Wylie, K. M., Wylie, T. N., Buller, R., Herter, B., Cannella, M. T., and Storch, G. A. (2018). Detection of viruses in clinical samples by use of metagenomic sequencing and targeted sequence capture. *J. Clin. Microbiol.* 56, e01123–18. doi: 10.1128/JCM.01123-18

Xu, B., Zhi, N., Hu, G., Wan, Z., Zheng, X., Liu, X., et al. (2013). Hybrid DNA virus in Chinese patients with seronegative hepatitis discovered by deep sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 110, 10264–10269. doi: 10.1073/pnas.1303744110

Xu, Y., Lewandowski, K., Lumley, S., Pullan, S., Vipond, R., Carroll, M., et al. (2018). Detection of viral pathogens with multiplex nanopore MinION sequencing: be careful with cross-talk. *Front. Microbiol.* 9:2225. doi: 10.3389/fmicb.2018.02225

Yang, J., Yang, F., Ren, L., Xiong, Z., Wu, Z., Dong, J., et al. (2011). Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *J. Clin. Microbiol.* 49, 3463–3469. doi: 10.1128/JCM.00273-11

Ye, S. H., Siddle, K. J., Park, D. J., and Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic classification. *Cell* 178, 779–794. doi: 10.1016/j.cell.2019.07.010

Yolken, R. H., Jones-Brando, L., Dunigan, D. D., Kannan, G., Dickerson, F., Severance, E., et al. (2014). Chlorovirus ATCV-1 is part of the human oropharyngeal virome and is associated with changes in cognitive functions in humans and mice. *Proc. Natl. Acad. Sci. U.S.A.* 111, 16106–16111. doi: 10.1073/pnas.1418895111

Yu, H. J., Deng, H., Ma, J., Huang, S. J., Yang, J. M., Huang, Y. F., et al. (2016). Clinical metagenomic analysis of bacterial communities in breast abscesses of granulomatous mastitis. *Int. J. Infect. Dis.* 53, 30–33. doi: 10.1016/j.ijid.2016.10.015

Zhao, G., Wu, G., Lim, E. S., Droit, L., Krishnamurthy, S., Barouch, D. H., et al. (2017). VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology* 503, 21–30. doi: 10.1016/j.virol.2017.01.005

Zheng, H., Jia, H., Shankar, A., Heneine, W., and Switzer, W. M. (2011). Detection of murine leukemia virus or mouse DNA in commercial RT-PCR reagents and human DNAs. *PLoS One* 6:e29050. doi: 10.1371/journal.pone.0029050

Zheng, Y., Gao, S., Padmanabhan, C., Li, R., Galvez, M., Gutierrez, D., et al. (2017). VirusDetect: an automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology* 500, 130–138. doi: 10.1016/j.virol.2016.10.017

Zhi, N., Hu, G., Wong, S., Zhao, K., Mao, Q., and Young, N. S. (2014). Reply to Naccache et al: viral sequences of NIH-CQV virus, a contamination of DNA extraction method. *Proc. Natl. Acad. Sci. U.S.A.* 111:E977. doi: 10.1073/pnas.1318965111

Zhong, Z.-P., Solonenko, N. E., Li, Y.-F., Gazitúa, M. C., Roux, S., Davis, M. E., et al. (2020). Glacier ice archives fifteen-thousand-year-old viruses. *bioRxiv*[Preprint]. doi: 10.1101/2020.01.03.894675

Zolfo, M., Pinto, F., Asnicar, F., Manghi, P., Tett, A., Bushman, F. D., et al. (2019). Detecting contamination in viromes using ViromeQC. *Nat. Biotechnol.* 37, 1408–1412. doi: 10.1038/s41587-019-0334-5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

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.