UCSF
UC San Francisco Previously Published Works

Title

The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns

Permalink

https://escholarship.org/uc/item/7qc1q1pb

Journal

Journal of Virology, 87(22)

ISSN

0022-538X

Authors

Naccache, Samia N
Greninger, Alexander L
Lee, Deanna
et al.

Publication Date

2013-11-15

DOI

10.1128/jvi.02323-13

Peer reviewed

The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns

Samia N. Naccache,[a,b] Alexander L. Greninger,[a,b] Deanna Lee,[a,b] Lark L. Coffey,[c] Tung Phan,[c] Annie Rein-Weston,[a,b] Andrew Aronsohn,[d] John Hackett, Jr.,[e] Eric L. Delwart,[a,c] Charles Y. Chiu[a,b,f]

Department of Laboratory Medicine, University of California, San Francisco, California, USA[a]; UCSF-Abbott Viral Diagnostics and Discovery Center, San Francisco, California, USA[b]; Blood Systems Research Institute, San Francisco, California, USA[c]; Center for Liver Disease, University of Chicago Medical Center, Chicago, Illinois, USA[d]; Abbott Diagnostics, Abbott Park, Illinois, USA[e]; Department of Medicine, Division of Infectious Diseases, University of California, San Francisco, California, USA[f]

Next-generation sequencing was used for discovery and *de novo* assembly of a novel, highly divergent DNA virus at the interface between the *Parvoviridae* and *Circoviridae*. The virus, provisionally named parvovirus-like hybrid virus (PHV), is nearly identical by sequence to another DNA virus, NIH-CQV, previously detected in Chinese patients with seronegative (non-A-E) hepatitis. Although we initially detected PHV in a wide range of clinical samples, with all strains sharing ~99% nucleotide and amino acid identity with each other and with NIH-CQV, the exact origin of the virus was eventually traced to contaminated silica-binding spin columns used for nucleic acid extraction. Definitive confirmation of the origin of PHV, and presumably NIH-CQV, was obtained by in-depth analyses of water eluted through contaminated spin columns. Analysis of environmental metagenome libraries detected PHV sequences in coastal marine waters of North America, suggesting that a potential association between PHV and diatoms (algae) that generate the silica matrix used in the spin columns may have resulted in inadvertent viral contamination during manufacture. The confirmation of PHV/NIH-CQV as laboratory reagent contaminants and not bona fide infectious agents of humans underscores the rigorous approach needed to establish the validity of new viral genomes discovered by next-generation sequencing.

Over the past 5 years, next-generation sequencing (NGS), otherwise known as deep sequencing, has been a remarkably successful approach for the identification and characterization of novel pathogens (1–3). In principle, with the exception of prions (4), all microbial agents that have the potential to cause disease can be detected in clinical samples on the basis of their specific nucleotide sequence. The rapidly increasing breadth and scope of microbial sequence reference databases in the research community have also facilitated the identification of novel microorganisms (3, 5), especially viruses that can exhibit a high degree of sequence divergence. Some recent examples of the use of NGS technology for pathogen discovery include identification of new agents associated with chronic illnesses, such as cancer (6, 7), screening of biologics, such as vaccines, for safety and testing purity (8–10), and outbreak investigation of novel viral pathogens (11–15).

Nevertheless, despite the broad utility of NGS in pathogen discovery, the technique is associated with a high risk of inadvertent contamination (16–18). The use of random instead of targeted primers to amplify all of the nucleic acid (NA) in clinical samples and the sheer depth of NGS, which routinely generates millions to billions of sequences per run, result in significant potential for laboratory and reagent contamination in addition to sample carryover. Concurrent analyses of cases and controls in a blinded fashion to exclude laboratory-derived contamination, collection of supportive clinical, epidemiologic, and serological data, and rigorous replication studies are thus critical in confirming or refuting putative associations of candidate novel agents with disease (3). These strategies were previously applied to conclusively determine that the retrovirus xenotropic murine leukemia virus-related virus(XMRV) is not associated with chronic fatigue syndrome or prostate cancer and, in fact, originated as a mouse cell line-derived laboratory contaminant (19–26).

Here we describe the identification and whole-genome assembly of a highly divergent single-stranded DNA (ssDNA) virus situated at the interface between *Circoviridae* and *Parvoviridae* by deep sequencing. The virus, provisionally named parvovirus-like hybrid virus (PHV), was detected in samples from patients with chronic seronegative (non-A-E) hepatitis and diarrhea of unknown etiology. The initial finding of a novel parvovirus-/circovirus-like agent in these patients was of great interest because these viruses are known to broadly infect insects, vertebrate animals, and humans (27–29), and specific members, such as parvovirus B19 in humans and porcine circovirus type 2 (PCV2) in pigs, have been linked to hepatitis (30–32). Furthermore, a study by Xu et al. recently described the discovery of a hybrid DNA virus in serum samples from Chinese patients with seronegative hepatitis, named NIH-CQV, with a sequence nearly identical to that of PHV (33). However, combined findings from follow-up deep sequencing, PCR, and data mining analyses performed in two independent laboratories and presented here demonstrate that PHV (and presumably NIH-CQV) are in fact laboratory reagent contaminants and underscore strategies that can be employed in the future to rapidly establish the significance and clinical relevance of novel microbial agents discovered by NGS.

MATERIALS AND METHODS

Ethics statement. All clinical samples were analyzed under protocols approved by the Institutional Review Boards (IRBs) of University of California, San Francisco (UCSF), and Blood Systems Research Institute (BSRI). Written informed consent was previously obtained for patients in non-A-E hepatitis cohort 1 (University of Chicago), non-A-E hepatitis cohort 2 (BIOLINCC Transfusion-Transmitted Viruses Study) [34], and the California Encephalitis Project [35] to provide clinical samples for viral analysis. Diarrheal samples from Nigeria, negative plasma for the HIV-spiked sample, and other miscellaneous clinical samples used to obtain the data shown in Fig. 2 and in Table S1 in the supplemental material did not require consent, as these samples were pre-existing and deidentified, and their use was thus deemed not to constitute human subject research.

Clinical sample sets corresponding to fully sequenced PHV genomes. (i) Non A-E hepatitis serum samples, cohort 1, United States. A total of 169 serum samples from a clinical study of patients with seronegative hepatitis were collected and processed at 3 different time points. In July 2010, 22 of these samples were extracted using the QIAamp UltraSens virus kit (Qiagen), and Illumina TruSeq-adapted libraries were constructed as described previously [13]. In February 2012, 64 samples, including the 22 previously processed samples, were combined in pools of 4, passed through a 0.22-μm filter (Millipore), and pretreated with nucleases as described previously [36], after which viral NA was isolated utilizing the QIAamp viral-RNA minikit (Qiagen), which purifies both RNA and DNA. Sixteen libraries spanning 64 samples total were constructed and deep sequenced across two Illumina HiSeq lanes (24 original samples per lane). In August of 2012, the remaining 105 samples were processed similarly to those processed in February 2012, but NA was isolated utilizing the EZ1 viral minikit v2.0 (Qiagen), and Illumina TruSeq-adapted libraries were prepared using an Eppendorf epMotion 5075 BioRobot.

(ii) Diarrheal stool samples, Nigeria. Seventy-five diarrheic stool samples from Nigeria and 50 from Tunisia were processed for viral NA extraction as previously described [37]. Briefly, samples were passed through a 0.45-μm filter and pretreated with nucleases, followed by NA extraction of 13 sample pools, each containing 5 to 10 samples, using the QIAamp viral-RNA minikit (Qiagen). Libraries were constructed using the ScriptSeq V2 RNA-Seq kit (Epicentre) [38] and deep sequenced on an Illumina MiSeq instrument.

(iii) Negative water control. A water sample was prepared as a negative control for a deep sequencing run. Briefly, the water sample was centrifuged at $12,000 \times g$ for 2 min and then passed through a 0.45-μm filter. Following prenuclease treatment, NAs were extracted using the QIAamp viral-RNA minikit (Qiagen). Libraries were then prepared using the ScriptSeq v2 RNA-Seq and run on an Illumina MiSeq instrument.

(iv) Non-A-E hepatitis serum samples, cohort 2, United States. A total of 88 serum samples from the Transfusion-Transmitted Virus Study (TTVS) [34] were obtained from the National Heart Lung and Blood Institute BioLINCC database (https://biolincc.nhlbi.nih.gov/studies/). In February of 2012, 16 serum samples from patients who developed hepatitis following transfusion were pooled in sets of 4 samples each and pretreated with nucleases as previously described [36]. Viral NA was isolated using the QIAamp viral-RNA kit (Qiagen), and 4 libraries were prepared using the ScriptSeq V2 RNA-Seq kit (Epicentre) and deep sequenced across one Illumina HiSeq lane. In August of 2012, all 88 individual samples were re-extracted following nuclease pretreatment using the EZ1 viral minikit v2.0 (Qiagen), and libraries were prepared using an epMotion 5075 BioRobot.

(v) HIV-1-spiked plasma sample. HIV-1 isolate DJ263 (CRF02_AG) was cultured, purified as described previously [39], quantified by the Abbott RealTime HIV-1 assay (Abbott Molecular, Des Plaines, IL), and used at a copy number of 10^4 to spike BaseMatrix defibrinated human plasma (Seracare) negative for hepatitis B virus (HBV), HCV, and HIV. Viral NA was extracted using the EZ1 viral minikit (Qiagen), followed by a post-DNase step and cleanup using the RNeasy MinElute cleanup kit (Qiagen). A TruSeq-adapted library was generated as above, followed by deep sequencing on an Illumina MiSeq instrument.

(vi) Clinical samples from the California Encephalitis Project. Cerebrospinal fluid, serum, lung swab, nasopharyngeal/throat swab, and brain tissue samples from patients with encephalitis were obtained as part of the California Encephalitis Project of the California Department of Public Health [35]. Samples were processed using the QIAamp viral-RNA minikit (Qiagen), and TruSeq-adapted libraries were generated as described above, followed by Illumina deep sequencing.

(vii) Miscellaneous clinical samples. Additional clinical samples screened for PHV include diarrheic stool samples from Mexico [40], *Naegleria fowleri* cultures [41], baboon adenovirus cultures [42], prostate cancer tissue [20], a hemorrhagic fever serum sample [13], human feces samples from Pakistan [37], and live-attenuated vaccines [8]. Unpublished viral metagenomic sequence data sets also analyzed by BLASTn for PHV consist of various human and animal plasma, tissue, respiratory fluid, and fecal samples, none of which included clinical samples from human hepatitis cases (additional details are available upon request) (see Table S1 in the supplemental material).

All NA extractions were performed according to the manufacturer's recommendations but excluded carrier RNA, which was replaced with linear acrylamide (Ambion). NGS libraries were quantified prior to deep sequencing as described previously [36] using a high-sensitivity DNA kit on an Agilent Bioanalyzer 2100 instrument (Agilent) and a KAPA library quantification kit (Kapa Biosystems).

PCR-based screening for PHV. To screen for PHV in various NA extraction kits, nuclease-free water from three sources (Fisher, Qiagen, and Epicentre) was mock extracted using the following kits according to the manufacturers' recommendations: RNeasy MinElute cleanup kit (Qiagen), RNeasy minikit (Qiagen), QIAamp UltraSens virus kit (Qiagen), QIAamp viral-RNA minikit (Qiagen), QIAamp DSP virus kit (Qiagen), PureLink viral-RNA/DNA minikit (Invitrogen), TRIzol reagent (Invitrogen), and EZ1 viral minikit v2.0 (Qiagen). To localize the source of PHV to the spin columns, the nuclease-free water was also directly eluted through the following spin columns: RNeasy MinElute spin columns (lot 136243239), RNeasy mini-spin columns (lot 1423422253), QIAamp mini-spin columns (lots 136267628, 136267629, and 139294474), QIAamp MinElute spin columns (lot 139305634), and PureLink viral-RNA/DNA spin columns. The water was eluted through at least three spin columns prior to PHV detection by PCR. Five fecal specimens from the diarrheal stool samples from Nigeria and five nuclease-free water samples were also re-extracted in parallel using the QIAamp viral-RNA minikit (Qiagen) and the MagMax viral-RNA isolation kit (Ambion) and tested for PHV by PCR.

PCR screening for PHV was performed using 10 μl of template and the HotStarTaq DNA polymerase kit (Qiagen) according to the manufacturer's suggestions. PCR amplicons were analyzed by gel electrophoresis and bands of the expected size were cloned in a TOPO vector (Invitrogen) and sequenced. Primer sets and PCR conditions for screening and confirmation of the PHV genome assembly are described in Table S2 in the supplemental material.

***De novo* assembly of PHV-1, PHV-2, and PHV-3 genomes.** Raw NGS reads from non-A-E hepatitis serum cohort 1, pool A (PHV-1), diarrheic stool from Nigeria, pool E (PHV-2), and extracted water (PHV-3) were preprocessed by quality filtering, primer trimming, and computational subtraction against human and bacterial reference databases to remove background as previously described [13], followed by BLASTx alignment to a viral protein database at an E-score cutoff of 10^{-3}. As some reads with significant homology to parvoviruses were detected, all computationally subtracted reads were then realigned to a targeted database of parvovirus proteins in GenBank using BLASTx at an E-score cutoff of 10^{-2}. Putative parvovirus reads were then inputted as "seeds" into the PRICE *de novo* assembler [43], requiring at least 85% identity over 25 nucleotides (nt) to merge two fragments. *De novo* assembly of the PHV genome was done

iteratively using PRICE and manual editing with the Geneious version 5.3.4 software package (44).

Screening for PHV in metagenomic data sets. Clinical NGS data sets generated in laboratories 1 and 2 (see Table S1 in the supplemental material) were screened for PHV by BLASTn alignment at E-score cutoffs of 10^{-30} (Illumina reads) or 10^{-50} (454 pyrosequencing contigs). 454 pyrosequencing reads were assembled into contigs using the SOAPdenovo package (45) prior to BLASTn alignment. Publicly available environmental metagenomic data sets deposited in the Sequence Read Archive (SRA), CAMERA (46), and MG-RAST (47) (see Table S3 in the supplemental material) databases were also screened by BLASTn alignment for PHV reads at an E-score cutoff of 10^{-30}.

Coverage maps. Reads from NGS data sets corresponding to *de novo*-assembled genomes PHV-1 (non-A-E hepatitis serum cohort 1, pool A) and PHV-2 (diarrheal stool, Nigeria, pool E) were aligned using BLASTn at a cutoff of 10^{-30} and mapped to their respective genomes using Geneious version 6.1.2. Coverage maps for other PHV strains were generated by BLASTn alignment at a cutoff of 10^{-30} and mapping to the PHV-1 genome. The consensus sequence was determined in Geneious by selection of the majority base at each nucleotide position.

Phylogenetic analysis. For construction of the amino acid phylogeny trees, the translated replicase and capsid sequences corresponding to representative parvoviruses, circoviruses, and circovirus-like viruses were first downloaded from GenBank (accession numbers are provided in the supplemental materials and methods). Multiple sequence alignments including PHV-1, PHV-2, and NIH-CQV were then performed using MAFFT with the "auto" option and with default parameters. A phylogenetic tree was constructed in Geneious version 6.1.2 with PHYML (48) with default parameters. Branch supports were computed in PHYML using an approximate likelihood ratio test (aLRT) approach based on an Shimodaira-Hasegawa-like (SH-like) option (49).

Nucleotide sequence accession numbers. The genome sequences of all 12 PHV strains described in this study (genotypes PHV-1, PHV-1B, PHV-1C, PHV-1D, PHV-2, PHV-3, PHV-4A, PHV-4B, PHV-5, PHV-6A, PHV-6B, and PHV-6C) have been deposited in GenBank as PHV strains UC1 to UC12 (accession numbers KF170373 and KF214637 to KF214647, respectively). The NGS data sets from which PHV-1, PHV-2, and PHV-3 were assembled were filtered for removal of human reads and deposited into the GenBank Sequence Read Archive (project accession number PRJNA217527 and SRA accession number SRP029352).

RESULTS

Discovery of PHV in seronegative hepatitis and diarrheal samples. As part of an ongoing investigation into potential viral etiologies for undiagnosed cases of seronegative non-A-E hepatitis, deep sequencing libraries in laboratory 1 at University of California, San Francisco (UCSF), were prepared from sera collected from a patient cohort of non-A-E hepatitis in the United States (Fig. 1A). Nucleic acids (NA) were extracted using the QIAamp viral-RNA minikit (Qiagen). Metagenomic libraries were prepared for unbiased deep sequencing from 64 patient samples which had been split into 16 indexed pools of 4 samples each. Analysis of the resulting NGS data using a previously developed computational pipeline for viral pathogen identification (17) revealed multiple divergent sequence reads with homology to parvoviruses. Translated amino acid alignments to parvovirus sequences in the GenBank nonredundant protein database (NR) resulted in the identification of a 100-bp read from one pool (pool A), sharing only 48% amino acid identity with *Acheta domestica* densovirus (ADZ50508.1, 93% query coverage, E value = 6×10^{-5}) and 39% identity with human parvovirus B19 (ABB36726.1, 93% query coverage, E value = 1×10^{-3}) in the capsid region (Fig. 1A, asterisk). This read was
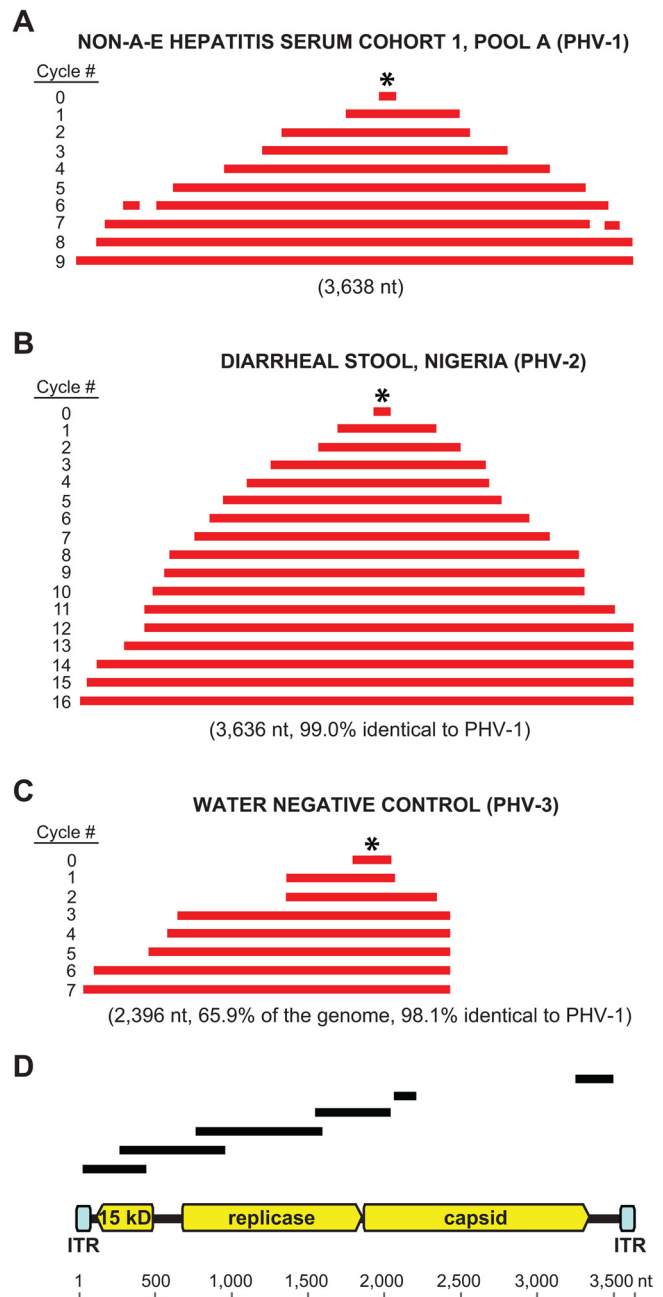

FIG 1 *De novo* assembly of PHV. The PRICE assembler was used to assemble the partial or full genomes of PHV-1 (A), PHV-2 (B), and PHV-3 (C), corresponding to a sample pool of non-A-E hepatitis sera, a sample pool of diarrheal stool samples, and a negative water control, respectively. The asterisks denote the initial seeds used for *de novo* assembly. Intermediate contiguous sequences (contigs) generated during the assembly (red bars) are mapped to their corresponding locations on the PHV genome. (D) Genome organization of PHV, showing the open reading frames (ORFs) corresponding to the putative replicase and capsid proteins and hypothetical 15-kDa protein of unknown function and two 148-nt inverted terminal repeat (ITR) sequences. Regions of the genome that were confirmed by Sanger sequencing are represented by black lines. nt, nucleotide.

selected as a seed for *de novo* assembly using the PRICE assembler (43), which generated a complete viral genome within 9 cycles (Fig. 1A). The organization of the viral genome was confirmed by PCR and Sanger sequencing of targeted regions

(Fig. 1D). The virus was provisionally named parvovirus-like hybrid virus (PHV), and the initial detected strain from the seronegative hepatitis pool was designated PHV-1.

A separate viral discovery lab, laboratory 2 at the Blood Systems Research Institute (BSRI), independently *de novo* assembled a novel parvovirus-like virus from NGS data corresponding to diarrheal stool samples from Nigeria (Fig. 1B). Nucleic acid extractions from these samples were also performed using the QIAamp viral-RNA minikit. *De novo* assembly of the viral genome was performed in 16 cycles from a single 100-bp read with 38% amino acid identity to Acheta domestica densovirus (ADZ50508.1, 96% query coverage, E value = 4×10^{-3}) and 39% identity to goose parvovirus (ABI20761.1, 84% query coverage, E value = 1×10^{-2}) (Fig. 1B, asterisk). Comparison of the assembled viral genome with PHV-1 revealed 99% nucleotide identity, and thus this virus was designated parvovirus-like hybrid virus, strain 2 (PHV-2). Strikingly, the whole-genome sequences of PHV-1 and PHV-2 shared 99% nucleotide and amino acid identity with each other and NIH-CQV, a novel hybrid DNA virus recently reported by Xu et al. to have been found in Chinese patients with seronegative hepatitis (33).

PHV genomic structure and phylogenetic analysis. The assembled genome of PHV-1 was found to be 3,636 bp long, with 3 open reading frames (ORFs) and 148-nt inverted terminal repeats at the 3′ and 5′ ends (Fig. 1D). The ORFs corresponding to the putative replicase and capsid genes were oriented in the same direction, and neither shared significant nucleotide identity with any sequence in NIH GenBank by BLASTn alignment. The replicase gene exhibited remote homology to circoviruses; by BLASTx protein alignment, ~25% of the translated sequence shared 35% amino acid identity with bat circovirus (AEL28794) and ~50% of the sequence shared 23% amino acid identity to porcine circovirus-like (Po-Circo-like) virus 21 (AER30018). The capsid gene was also highly divergent, with only 31% translated amino acid identity over 17% of the gene to the corresponding capsid gene in goose parvovirus (ACK86566.1). Sequences encoding a conserved P-loop nucleoside triphosphate (NTP)-binding domain (50) and N-terminal parvovirus coat domain (51) were detected in the replicase and capsid genes, respectively. The capsid gene also encoded a putative phospholipase A$_2$ (PLA$_2$) motif that is critical for parvovirus infectivity (52). Bridging PCR using primers spanning the two largest ORFs confirmed that the circovirus-like replicase and parvovirus-like capsid genes originated from the same viral genome (Fig. 1D). Multiple attempts using inverse PCR failed to detect evidence of a circular form for PHV. To determine the exact phylogenetic placement of PHV relative to other ssDNA viruses, amino acid phylogenetic analysis of the putative replicase and capsid proteins was performed using representative genomes from the families *Circoviridae* and *Parvoviridae*. The resulting phylogenetic trees (Fig. 2) revealed that PHV and NIH-CQV are situated on a deep independent branch that appears to be intermediate between the circoviruses and parvoviruses. The closest, albeit distant, relative to PHV and NIH-CQV is Po-Circo-like virus 21, a porcine circovirus-like virus previously identified as part of the fecal virome of pigs at a high-density farm (53).

Failure to detect PHV in re-extracted clinical samples. In non-A-E hepatitis serum cohort 1, the observation that reads from PHV-1 were present in all of the indexed pools (Fig. 3A) raised the likelihood of contamination, either laboratory derived or from sample cross-contamination. To investigate this possibility, indi-

vidual samples corresponding to the pool from which the PHV-1 genome was initially assembled (pool A) were re-extracted using a magnetic-bead-based NA extraction method on an automated EZ1 instrument (EZ1 viral minikit v2.0) and tested for PHV by specific PCR. Although PCR of the NA extracted using the QIAamp viral-RNA minikit successfully detected PHV in all of the tested samples, NA extracted from the same samples using the automated instrument tested negative for PHV. These discrepant results raised doubts as to whether the original clinical samples actually harbored PHV.

Detection and genome assembly of PHV from multiple sample cohorts. To further investigate the prevalence of PHV in clinical samples, BLASTn alignments of 28 metagenomic data sets corresponding to a wide range of clinical sample cohorts were performed, using a high-stringency E value of 10^{-30} (Illumina 100-bp or 250-bp short reads) or 10^{-50} (longer Roche 454 pyrosequencing reads) for detection of PHV sequences. In non-A-E hepatitis serum cohort 1, reads aligning to PHV were detected in all pools, with the percentage of total reads per pool being remarkably similar, between 0.2 and 0.3% (Fig. 3A). PHV sequences were identified in multiple additional data sets from laboratories 1 and 2 (Fig. 3A and B). Sample data sets positive for PHV had all been processed using a column-based Qiagen NA extraction kit (Fig. 3A and B, red text), with most of the detected PHV sequences corresponding to samples processed from 2011 to the present. No PHV reads were associated with data sets corresponding to NA isolated using kits from manufacturers other than Qiagen or by other extraction methods (e.g., use of magnetic beads) (Fig. 3A and B, black text). In overlapping samples from non-A-E hepatitis serum cohorts 1 and 2 that had been extracted using two independent methods (Fig. 3A), i.e., by using Qiagen column-based and magnetic bead-based kits, PHV reads were detected only in samples that had been extracted using Qiagen columns (Fig. 3A). Strikingly, a large number of PHV reads were also recovered from deep sequencing of a negative water control mock extracted through the QIAamp viral-RNA minikit, from which ~2/3 of the genome, designated PHV-3, could be *de novo* assembled (Fig. 1C).

The average coverage of PHV-1 and PHV-2 achieved by deep sequencing of the sample pools and *de novo* assembly was 993× and 195×, respectively, and spanned 99 to 100% of the genome (Fig. 4A). Although the complete PHV-3 genome derived from the negative water control could not be *de novo* assembled due to a gap in coverage (Fig. 4A, arrow), the PHV-3 reads mapped to PHV-1 spanned 97% of the genome at 69× coverage. The average coverage obtained from representative individually indexed samples or sample pools when mapped to PHV-1 was >150× and spanned 98 to 100% of the genome (Fig. 4B). Consensus sequences generated from all 12 coverage maps revealed that all assembled viral genomes were remarkably similar (Fig. 5A), diverging from the PHV-1 reference strain by <1.3%, with the exception of PHV-3 (negative water control), which diverged by 4.2% due to gaps in coverage. Notably, all PHV consensus sequences were found to share 99 to 100% amino acid identity with each other and with NIH-CQV.

PHV is a viral contaminant of Qiagen spin columns. The detection of PHV reads only in metagenomic data sets corresponding to clinical samples extracted using Qiagen spin columns (Fig. 3A and B), and the assembly of a PHV genome directly from a negative water control that had also been processed using the same spin columns (Fig. 1C), raised the strong possibility that Qiagen
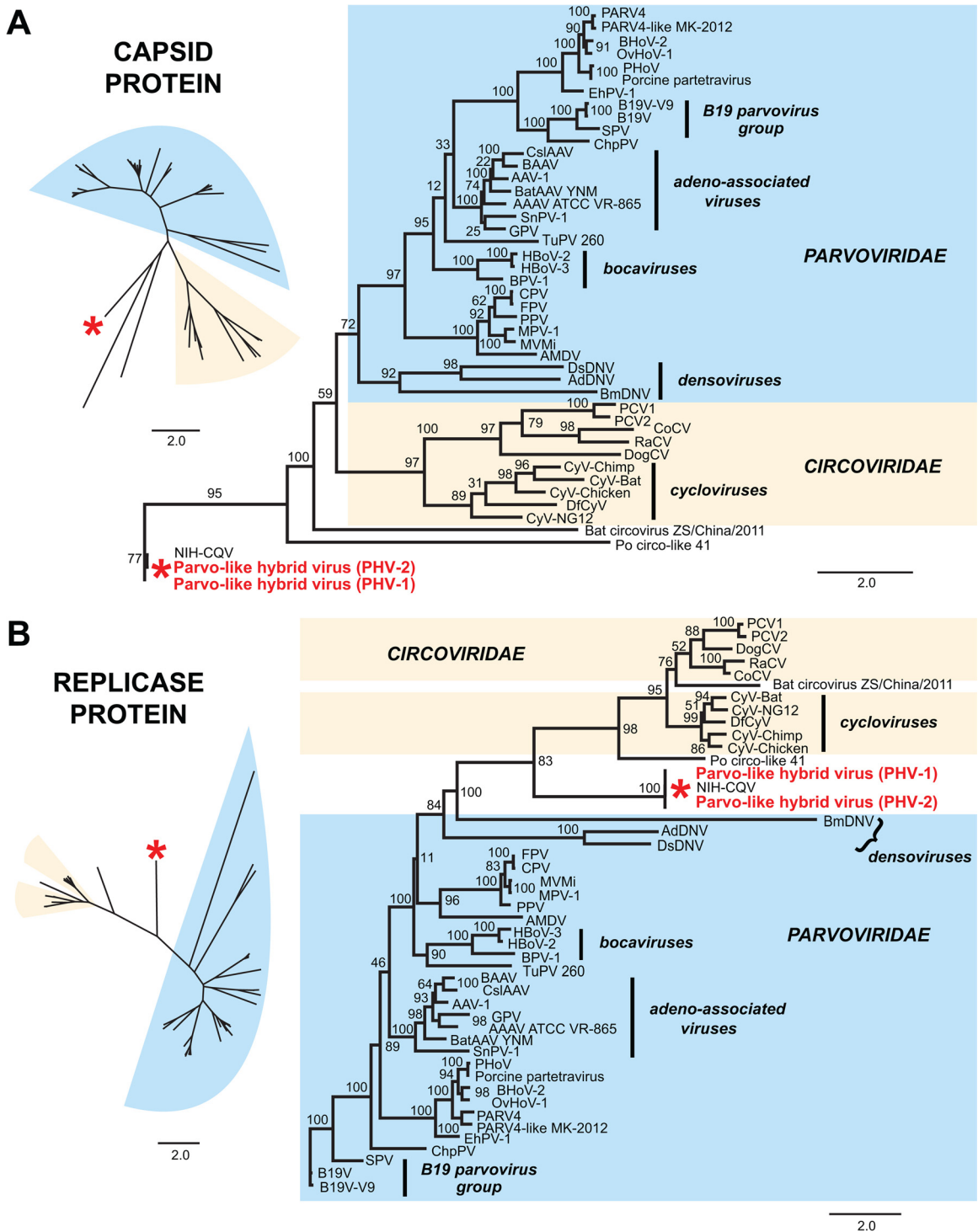
FIG 2 Amino acid phylogenetic trees of PHV-1, PHV-2, and NIH-CQV relative to other ssDNA viruses. (A) Capsid protein; (B), replicase protein. Representative parvoviruses, circoviruses, and circovirus-like viruses were included in the phylogenetic analysis. Support levels calculated using a maximum likelihood algorithm are displayed at each branching point. Scale bars indicate the number of amino acid substitutions per site. Abbreviations and GenBank accession numbers are reported in the supplemental materials and methods.

columns were contaminated with PHV. To confirm this suspicion, mock extractions of water were performed using a variety of spin columns from Qiagen, spin columns from a different manufacturer (Invitrogen), magnetic bead-based extraction kits (Qiagen EZ1 and Ambion), and TRIzol (Invitrogen) (Table 1; also see Fig. S1B in the supplemental material). Direct mock elutions of water through the spin columns were also performed (see Fig. S1A in the supplemental material). PCR analysis of the processed water
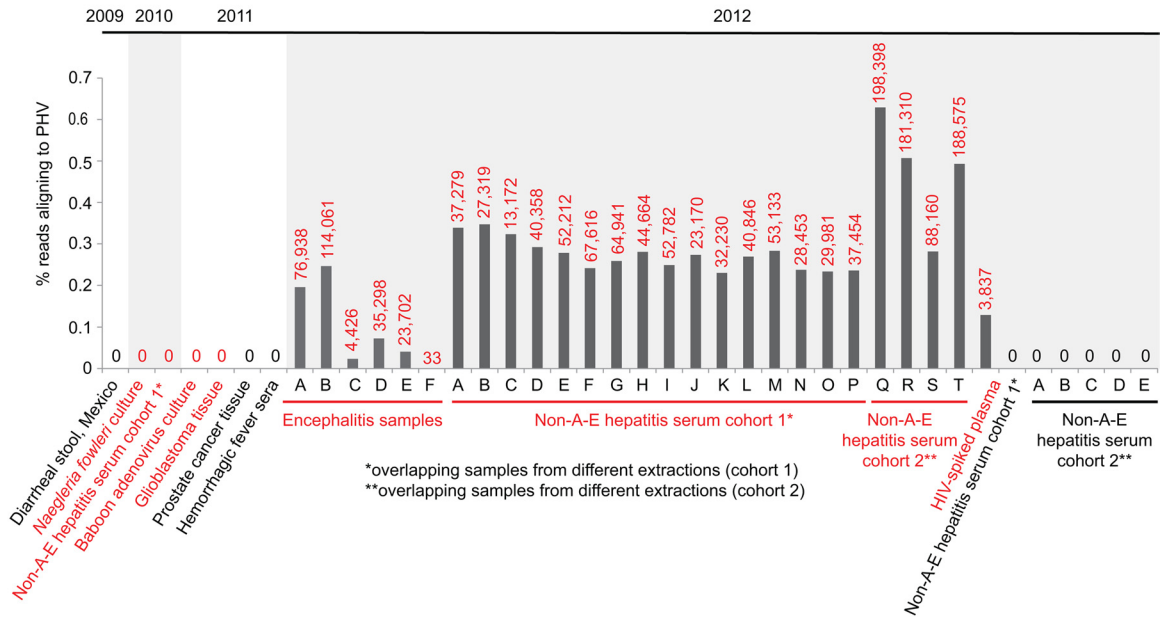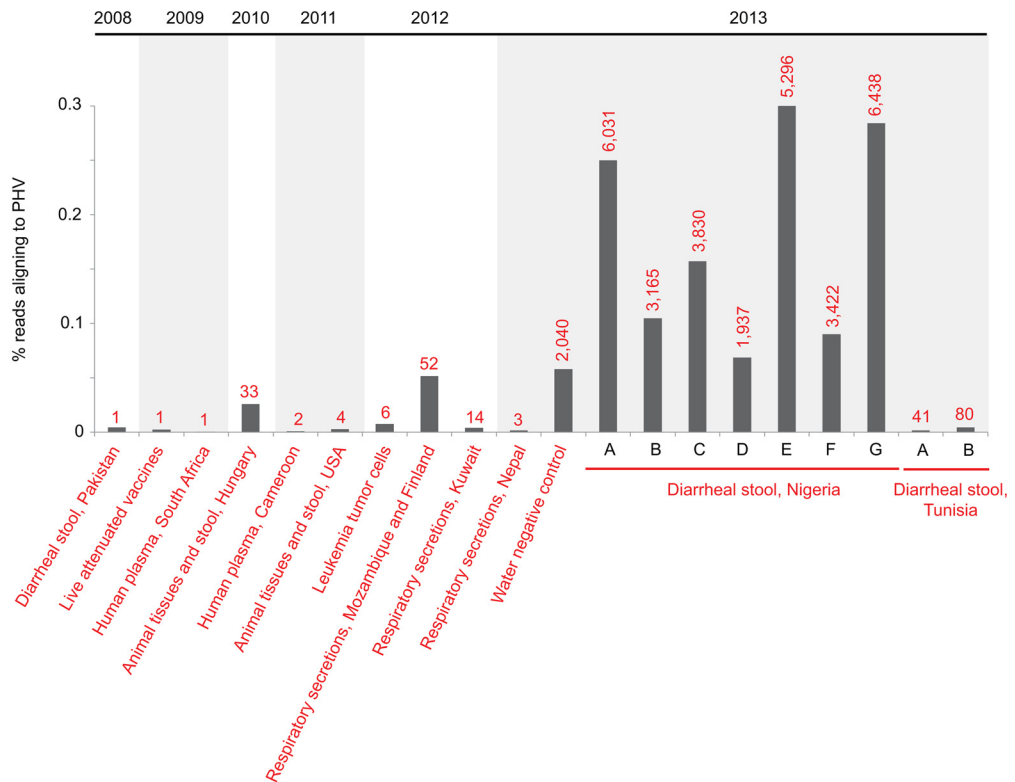
FIG 3 Screening for PHV in clinical NGS data sets. Sequences corresponding to PHV were identified by BLASTn alignments at an E-score cutoff of 10^{-30} to sequence reads in NGS data sets processed from 2008 to 2013 in laboratory 1 (A) and laboratory 2 (B). The bar graphs show the percentage of PHV reads in each data set, with the number of PHV reads given above the bars. Clinical sample sets extracted using Qiagen spin columns are highlighted using red text; those extracted using other methods are shown in black. Abbreviations: UCSF, University of California, San Francisco; BSRI, Blood Systems Research Institute.
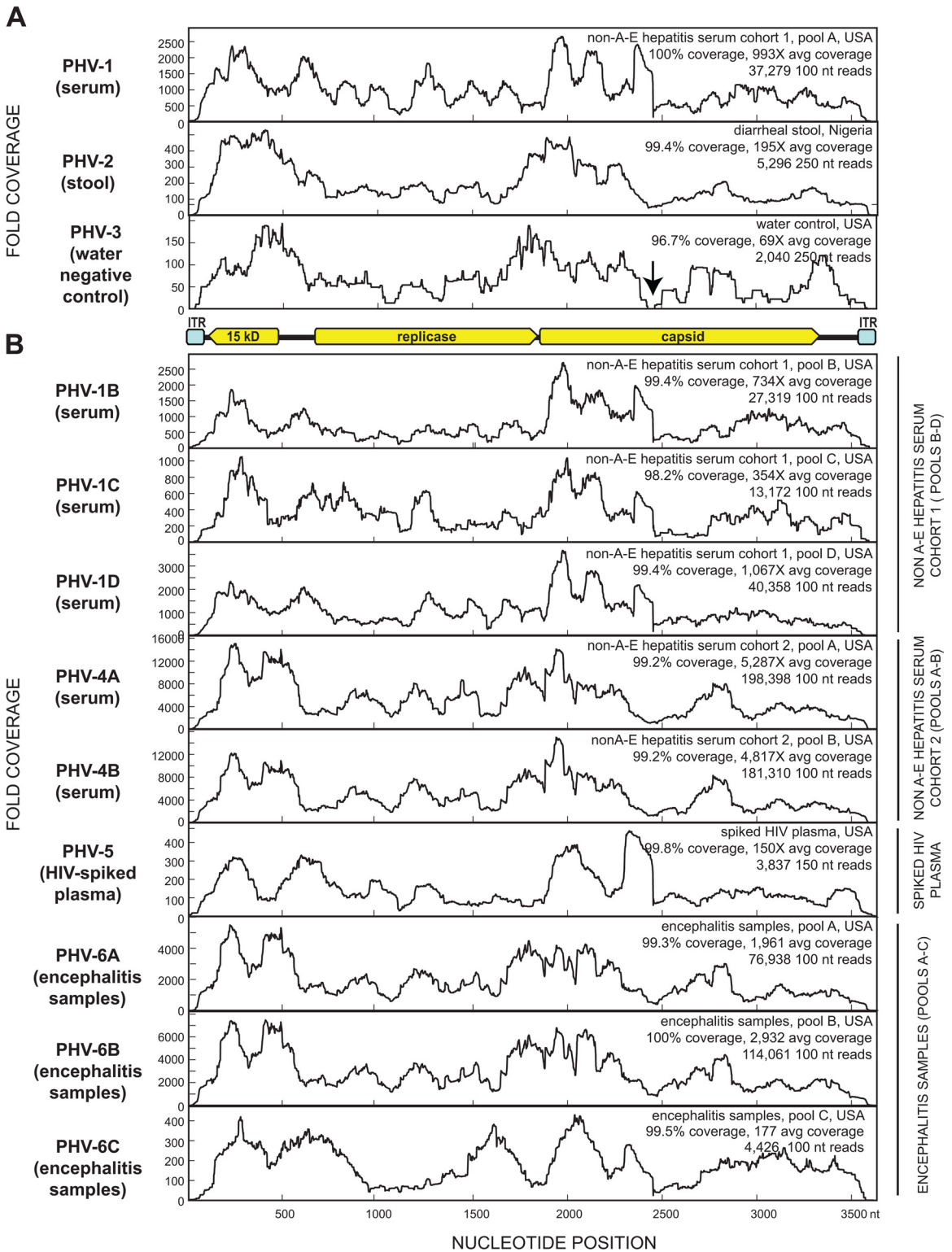
FIG 4 Mapping of reads from clinical NGS data sets to the PHV genome. The coverage (*y* axis) achieved at each nucleotide position along the genome (*x* axis) is plotted. (A) Coverage maps for the independently *de novo* assembled PHV-1, PHV-2, and PHV-3 genomes. The arrow depicts a break in coverage that precluded whole-genome *de novo* assembly of PHV-3. (B) PHV genomes assembled by mapping of NGS reads from various clinical sample data sets to the PHV-1 genome. The inset information includes the name of the clinical data set, as well as the percent (%) and average (avg) genomic coverage achieved. nt, nucleotide.
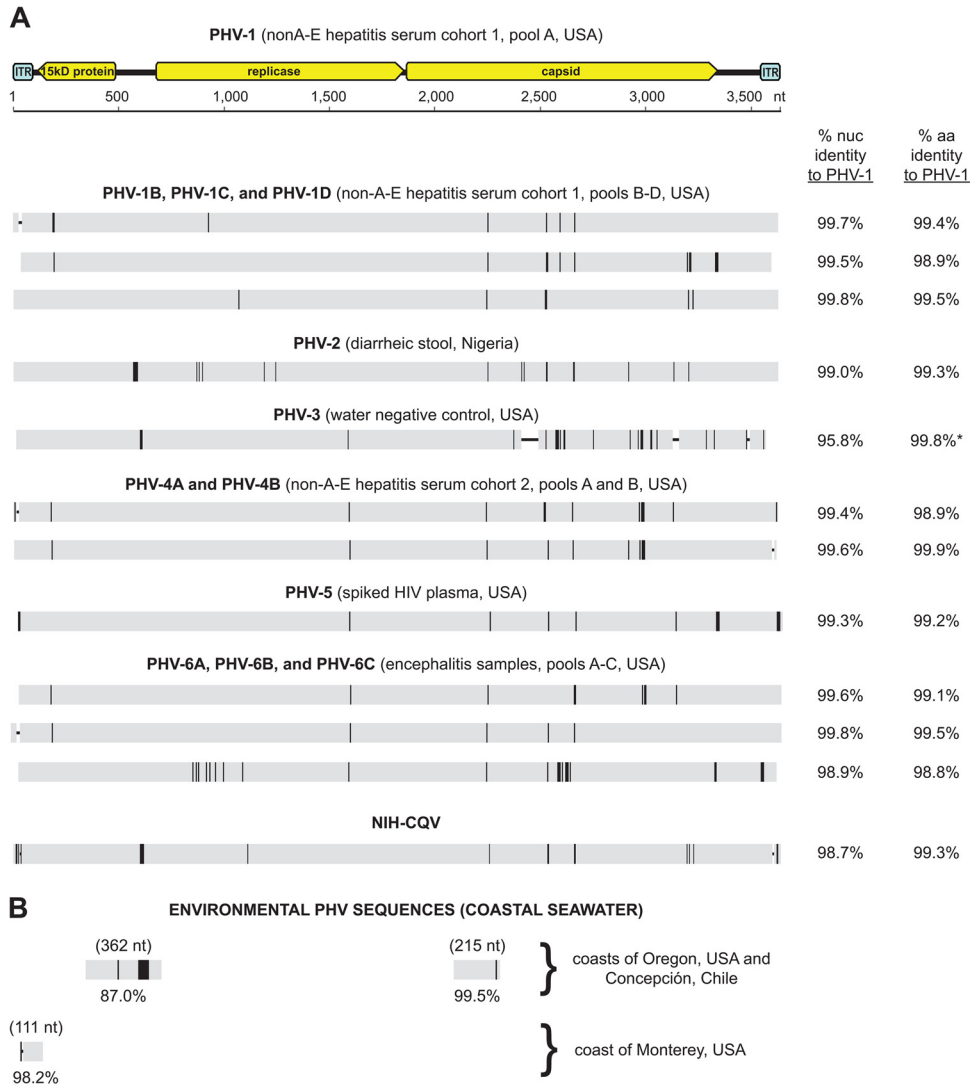
FIG 5 Whole-genome pairwise comparison of PHV strains and NIH-CQV relative to PHV-1. Black vertical bars indicate mismatches or insertions, while horizontal lines indicate deletions in the whole-genome consensus sequence relative to PHV-1. nt, nucleotide.

controls using four sets of primers detected PHV only in samples that had been processed using Qiagen spin columns, while using spin columns from other manufacturers or different extraction methods consistently failed to detect PHV. A subset of PCR amplicons were Sanger sequenced and confirmed to be >99% identical to PHV-1. The detection of PHV in mock water extractions through Qiagen columns was reproducible in two independent laboratories (see Fig. S2A in the supplemental material) and with the use of purified water from multiple sources (Fisher, Qiagen, and Epicentre) (see Fig. S2B), directly implicating Qiagen spin columns as the source of PHV contamination.

Identification of PHV sequences in environmental samples. To gain further insight into the origins of PHV, publicly available environmental metagenomic sequence data sets in the CAMERA (46) and MG-RAST (47) databases were scanned for evidence of PHV-related sequences using BLASTn alignments at a high-stringency cutoff of 10^{-30}. A total of 78 public data sets containing 213,615,095 sequence reads were analyzed, including 8,063,303 reads from vertebrate metagenomes, 395,038 reads from plant

metagenomes, 14,922,577 reads from sediment sewage and soil metagenomes, 6,609,658 reads from freshwater metagenomes, and 189,242,666 reads from marine metagenomes, including plankton, microbialite, and coral reef metagenomic studies. Two NGS data sets from marine sources containing Roche 454 pyrosequencing data were found to harbor 3 PHV sequences, in total spanning 17% of the PHV-1 genome with 87 to 99% identity (Fig. 5C). Interestingly, both data sets corresponded to metagenomic shotgun sequencing of sampled seawater off the Pacific coast of North America, as two of the identified PHV reads were derived from a study of metagenomes in Monterey Bay, California (CAMERA project "North Pacific metagenomes from Monterey Bay to Open Ocean") (46, 54), and one was from a study of metagenomes in coastal regions off Oregon and Concepción, Chile (CAMERA project "Microbial initiative in low oxygen areas off Concepción and Oregon") (see Table S3 in the supplemental material) (46). As the sample processing and generation of these data sets did not involve the use of columns or kits manufactured by Qiagen (A. Z. Worden, A. Bertagnolli, and S. Giovannoni, personal communi-

TABLE 1 PCR screening of commonly used viral nucleic acid extraction kits for parvovirus-like hybrid virus (PHV-1)[a]

Kit	Spin column	Replicase, nt763-1010 (248 nt)		Bridge, nt1554-2044 (491 nt)		Capsid, nt1922-2044 (121 nt)		Capsid + NCR, nt3288-3448 (161 nt)	
		C	F	C	F	C	F	C	F
RNeasy MinElute cleanup kit	RNeasy MinElute column	+	+	−	+	+	+	+	+
RNeasy minikit	RNeasy minicolumn	+	+	+	+	+	+	+	+
QIAamp UltraSens virus kit	QIAamp minicolumn	+	+	−	+	+	+	+	+
QIAamp viral RNA minikit	QIAamp minicolumn	−	+	−	−	+	+	+	+
QIAamp DSP virus kit	QIAamp MinElute column	−	+	−	−	-	+	-	+
PureLink viral RNA/DNA minikit	PureLink viral column	−	−	−	−	-	-	-	-
TRIzol LS kit	NA	−	−	−	−	-	-	-	-
EZ1 viral minikit v2.0	NA	−	−	−	−	-	-	-	-
Water, nuclease-free (Qiagen, Fisher Scientific, and Epicentre)	NA	−	−	−	−	-	-	-	-

[a] NCR, noncoding region; C, column elution; F, full extraction; nt, nucleotide; NA, not applicable.

cation), these findings revealed that PHV is an environmental virus likely originating in ocean water and may have inadvertently contaminated the spin columns during manufacture.

DISCUSSION

Here we present the discovery of a novel, highly divergent DNA virus, PHV, that appears to be a hybrid intermediate between the parvoviruses and circoviruses and is nearly identical by sequence to another virus, NIH-CQV, recently reported to have been found in patients with seronegative hepatitis. The arrangement of the linear ~3.6-kb genome of PHV, consisting of two forward-direction ORFs encoding the replication and capsid proteins and a pair of inverted terminal repeats, is characteristic of viruses in *Parvoviridae*, and PHV retains the P-loop NTP-binding domain in the replicase protein and PLA$_2$-like and parvovirus coat motifs in the capsid protein that are conserved among members of that family. Although the replicase gene of PHV is more similar to that of circoviruses by direct BLASTx searches, we were unable to recover a closed circular form of PHV by inverse PCR, unlike the previous study by Xu et al. describing the discovery of the related NIH-CQV (33). If present in PHV, an episomal form, which has been detected in parvoviruses such as adeno-associated virus and bocavirus (55, 56), may represent a rare variant in the viral population. By phylogenetic analysis, PHV does appear to bridge the interface between circoviruses and parvoviruses (Fig. 2) and, as such, may be related to the last common ancestor of these two ssDNA viral family lineages.

The genome of PHV was independently detected and *de novo* assembled in two laboratories from pooled samples from patients with non-A-E hepatitis or diarrhea. However, subsequent identification of PHV sequences in multiple, widely different sample sets and failure to confirm the presence of PHV after re-extraction of NA by a method other than the use of Qiagen spin columns raised a strong suspicion that PHV may be a laboratory contaminant. This was directly confirmed by PCR and deep sequencing of extracted and eluted water controls, revealing that PHV-specific sequences could be directly recovered from Qiagen spin columns. The degree of potential contamination by PHV is significant given

that a contiguous stretch comprising ~2/3 of the genome could be *de novo* assembled from deep sequencing data corresponding to a negative water control (Fig. 1C). Even more strikingly, direct mapping of NGS reads corresponding to the water control to the PHV-1 reference sequence enabled assembly of >97% of the genome (Fig. 3A). The high efficiency of silica-based spin columns in concentrating DNA/RNA during extraction (57) may have played a role in amplifying even trace contamination from PHV occurring during the time of manufacture. Contamination was observed only in spin columns from a single manufacturer (Qiagen) and was not seen with columns from other manufacturers (Invitrogen), or with alternative methods of extraction, such as TRIzol (Invitrogen) or magnetic beads (Ambion or Qiagen EZ1). Furthermore, the extent of contamination appeared to be time and/or batch dependent, as spin column-based kits manufactured prior to 2011 were largely devoid of PHV sequences but those made in 2012 and 2013 were likely to be heavily contaminated (Fig. 3A and B). Such sporadic contamination events may inadvertently mislead researchers into erroneously making disease associations if they are unaware that a newly discovered virus is a contaminant and not a bona fide infectious agent. The presence of contamination in spin columns, such as the previously reported detection of sequences corresponding to murine DNA, circoviruses/densoviruses, and *Legionella* bacteria in Qiagen NA extraction columns (16, 58, 59), can also negatively impact the performance of both clinical and research-based assays for pathogen detection, underscoring the need for DNA-free reagents.

In the present study, the consensus genomes assembled from the various PHV strains were remarkably similar, exhibiting 96 to 100% nucleotide identity with each other (Fig. 4). The very slight observed differences may reflect natural variation or errors in the deep sequencing, either native to the technology or due to sequencing artifacts from random priming or PCR duplication. Strikingly, on an amino acid level, the translated sequences for the major proteins were 99 to 100% identical across all 12 assembled PHV genomes and NIH-CQV. Our finding of very low intrastrain variation in the PHV genome contrasts markedly with that described by Xu et al., in which significant genetic heterogeneity in

NIH-CQV corresponding to putative sequence variants between patient samples was observed (33). Although fold coverage maps from that prior study were not presented, it is possible that insufficient sequence coverage and/or errors in the NGS data may have accounted for the observed high substitution rates. Notably, in our study, greater sequence variation in the assembled genomes was observed in PHV-3 (negative water control) and PHV-6C (encephalitis samples, pool C), which had comparatively lower depths of coverage than the other PHV genomes (Fig. 3 and 5). An alternative possibility is that there is indeed genetic heterogeneity in PHV/NIH-CQV that reflects natural variation and/or artifactual variation arising from lot-to-lot variability in the degree of spin column contamination.

The finding of laboratory contamination as the origin of PHV suggests that NIH-CQV, which shares 100% amino acid identity with PHV, is most likely also a laboratory contaminant. In the study by Xu et al., there was 70% PCR positivity in seronegative hepatitis patient samples with an average virus titer of 1.05×10^4 copies/μl (corresponding to 1.05×10^7 copies/ml) yet 0% positivity in healthy blood donors (33). The dichotomy between these results and serological detection showing comparable rates of positivity for IgG specific to the C-terminal portion of the NIH-CQV capsid protein in hepatitis patients and blood donors is striking. The PCR results may be explained by lot-to-lot variability or the use of a Qiagen extraction kit prior to 2011, as those kits appeared to be less contaminated with PHV/NIH-CQV, while the serological results may potentially be due to detection of cross-reactive antibodies by the immunoblot assay. Previously, a serological assay designed to detect antibodies to p15E of XMRV showed elevated seroreactivity in human T-cell lymphotropic virus type 1 (HTLV-1)-infected individuals (60), although none of these individuals had detectable antibodies to a second XMRV protein, gp70. Subsequent analysis revealed that a highly conserved sequence within the immunodominant region of HTLV gp21 that is shared with p15E was likely the source of the cross-reactive antibodies elicited by HTLV-1 infection (60). In the study by Xu et al. (33), confirmatory data based on serologic reactivity to multiple nonoverlapping epitopes within a single protein or more than one viral protein would have provided stronger evidence of infection by NIH-CQV.

By data mining of publicly available environmental metagenomic databases, sequences with 100% identity to PHV/NIH-CQV were detected in coastal waters off North America. The relatively low number of reads detected is likely due to several factors: (i) high-efficiency concentration of viral DNA in the spin columns, (ii) differential rates of PHV abundance in ocean water, and (iii) lower-throughput Roche 454 pyrosequencing rather than Illumina NGS for data generation. Viral abundance in aquatic ecosystems is exceedingly high, with concentrations estimated at $\sim 10^8$ per 1 ml (61). In total, approximately 10^{30} viruses are thought to reside in the world's oceans, constituting a vast, largely unsequenced reservoir of genomes. In addition, highly diverse ssDNA viruses, such as circoviruses and parvoviruses, have been detected in seawater (62) and in ocean dwellers such as peneid shrimp (63), and viruses are known to infect diatoms (algae) that are ubiquitous in seawater (64, 65). Taken together, these observations suggest a plausible pathway for how PHV contamination of the NA spin columns could have occurred. Column-based NA purification is a solid-phase extraction method that binds NA by adsorption to silica, and the silica used in many commercial spin columns is derived from the cell walls of diatoms (57). If Qiagen's NA extraction kits and "silica gel membrane technology" involve the use of diatoms (66), it is plausible that PHV is a virus of diatoms and had inadvertently contaminated the spin columns during manufacture. The sporadic contamination observed in the silica-based spin columns (Fig. 3A and B) may thus be due to seasonal variation in diatom abundance, diatom type, and rates of viral infection (67). The contamination of spin columns is not confined to PHV but can also be seen by the presence of sequences corresponding to phages, circoviruses, and parvoviruses other than PHV (16). Further studies will be needed to establish that PHV is a virus of diatoms. Notably, we did not detect PHV in environmental metagenomic data sets corresponding to other oceanic or environmental communities, which may reflect a limited geographic and temporal distribution for the virus or a bias and/or incompleteness in the publicly available metagenomic databases surveyed. The impact, if any, of these oceanic viruses on human health or public safety is unknown.

As the use of molecular methods such as deep sequencing for pathogen discovery becomes more frequent, it is critical that robust strategies be developed to rapidly determine the biological and clinical relevance of any new candidate agent. This is especially true with the discovery of novel, potentially transfusion-transmissible viruses in blood that may have an immediate impact on infectious diseases and public health (68), as exemplified by the high-profile putative association between XMRV and chronic fatigue syndrome that was eventually refuted by rigorous follow-up investigation (19–26). In the present study, the confirmation of PHV as a laboratory reagent contaminant and not a candidate blood-borne infectious agent was made possible by (i) independent assessment at two research sites, (ii) free and open sharing of sequence data corresponding to multiple sample cohorts between laboratories, (iii) use of control samples subjected to the same extraction and deep sequencing steps as experimental samples, (iv) direct PCR confirmation of viral contamination, and (v) data mining of publicly available metagenomic sequence databases derived from a vast array of clinical and environmental samples. Our results thus strongly call into question any association of the PHV and NIH-CQV viruses with seronegative hepatitis or, indeed, any bona fide infections of humans. Timely reporting of "dediscoveries" as well as discoveries, by focusing effort and resource investment, is needed to maximize the translational impact of pathogen discovery to clinical medicine and infectious diseases.

ACKNOWLEDGMENTS

REFERENCES

1. **Chiu CY.** 2013. Viral pathogen discovery. Curr. Opin. Microbiol. http://dx.doi.org/10.1016/j.mib.2013.05.001. [E-pub ahead of print.]
2. **Lipkin WI.** 2013. The changing face of pathogen discovery and surveillance. Nat. Rev. Microbiol. **11:**133–141.

3. **Lipkin WI.** 2010. Microbe hunting. Microbiol. Mol. Biol. Rev. **74:**363–377.

4. **Colby DW, Prusiner SB.** 2011. Prions. Cold Spring Harb. Perspect. Biol. **3:**a006833. doi:10.1101/cshperspect.a006833.

5. **Relman DA.** 1998. Detection and identification of previously unrecognized microbial pathogens. Emerg. Infect. Dis. **4:**382–389.

6. **Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J, Baselga J, Liu C, Shivdasani RA, Ogino S, Birren BW, Huttenhower C, Garrett WS, Meyerson M.** 2012. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. **22:**292–298.

7. **Feng H, Shuda M, Chang Y, Moore PS.** 2008. Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science **319:**1096–1100.

8. **Victoria JG, Wang C, Jones MS, Jaing C, McLoughlin K, Gardner S, Delwart EL.** 2010. Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus. J. Virol. **84:**6033–6040.

9. **Chiu C.** 2013. The hidden threat of unidentified agents of disease in human and veterinary biologicals. Biologicals **41:**129–130.

10. **Chandriani S, Skewes-Cox P, Zhong W, Ganem DE, Divers TJ, Van Blaricum AJ, Tennant BC, Kistler AL.** 2013. Identification of a previously undescribed divergent virus from the Flaviviridae family in an outbreak of equine serum hepatitis. Proc. Natl. Acad. Sci. U. S. A. **110:**E1407–E1415.

11. **Xu B, Liu L, Huang X, Ma H, Zhang Y, Du Y, Wang P, Tang X, Wang H, Kang K, Zhang S, Zhao G, Wu W, Yang Y, Chen H, Mu F, Chen W.** 2011. Metagenomic analysis of fever, thrombocytopenia and leukopenia syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus. PLoS Pathog. **7:**e1002369. doi:10.1371/journal.ppat.1002369.

12. **Palacios G, Druce J, Du L, Tran T, Birch C, Briese T, Conlan S, Quan PL, Hui J, Marshall J, Simons JF, Egholm M, Paddock CD, Shieh WJ, Goldsmith CS, Zaki SR, Catton M, Lipkin WI.** 2008. A new arenavirus in a cluster of fatal transplant-associated diseases. N. Engl. J. Med. **358:**991–998.

13. **Grard G, Fair JN, Lee D, Slikas E, Steffen I, Muyembe J-J, Sittler T, Veeraraghavan N, Ruby G, Wang C, Makuwa M, Mulembakani P, Mazet J, Rimoin A, Taylor T, Schneider BS, Simmons G, Delwart E, Wolfe ND, Chiu CY, Leroy E.** 2012. A novel rhabdovirus associated with acute hemorrhagic fever in Central Africa. PLoS Pathog. **8:**e1002924. doi:10.1371/journal.ppat.1002924.

14. **Briese T, Paweska JT, McMullan LK, Hutchison SK, Street C, Palacios G, Khristova ML, Weyer J, Swanepoel R, Egholm M, Nichol ST, Lipkin WI.** 2009. Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. PLoS Pathog. **5:**e1000455. doi:10.1371/journal.ppat.1000455.

15. **Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA.** 2012. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N. Engl. J. Med. **367:**1814–1820.

16. **Lysholm F, Wetterbom A, Lindau C, Darban H, Bjerkner A, Fahlander K, Lindberg AM, Persson B, Allander T, Andersson B.** 2012. Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. PLoS One **7:**e30875. doi:10.1371/journal.pone.0030875.

17. **Greninger AL, Chen EC, Sittler T, Scheinerman A, Roubinian N, Yu G, Kim E, Pillai DR, Guyard C, Mazzulli T, Isa P, Arias CF, Hackett J, Schochetman G, Miller S, Tang P, Chiu CY.** 2010. A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. PLoS One **5:**e13381. doi:10.1371/journal.pone.0013381.

18. **Cheval J, Sauvage V, Frangeul L, Dacheux L, Guigon G, Dumey N, Pariente K, Rousseaux C, Dorange F, Berthet N, Brisse S, Moszer I, Bourhy H, Manuguerra CJ, Lecuit M, Burguiere A, Caro V, Eloit M.** 2011. Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. J. Clin. Microbiol. **49:**3268–3275.

19. **Simmons G, Glynn SA, Komaroff AL, Mikovits JA, Tobler LH, Hackett J, Jr, Tang N, Switzer WM, Heneine W, Hewlett IK, Zhao J, Lo SC, Alter HJ, Linnen JM, Gao K, Coffin JM, Kearney MF, Ruscetti FW, Pfost MA, Bethel J, Kleinman S, Holmberg JA, Busch MP, Blood XSRWG.** 2011. Failure to confirm XMRV/MLVs in the blood of patients with chronic fatigue syndrome: a multi-laboratory study. Science **334:**814–817.

20. **Lee D, Das Gupta J, Gaughan C, Steffen I, Tang N, Luk KC, Qiu X, Urisman A, Fischer N, Molinaro R, Broz M, Schochetman G, Klein EA, Ganem D, Derisi JL, Simmons G, Hackett J, Jr, Silverman RH, Chiu CY.** 2012. In-depth investigation of archival and prospectively collected samples reveals no evidence for XMRV infection in prostate cancer. PLoS One **7:**e44954. doi:10.1371/journal.pone.0044954.

21. **Knox K, Carrigan D, Simmons G, Teque F, Zhou Y, Hackett J, Jr, Qiu X, Luk KC, Schochetman G, Knox A, Kogelnik AM, Levy JA.** 2011. No evidence of murine-like gammaretroviruses in CFS patients previously identified as XMRV-infected. Science **333:**94–97.

22. **Alter HJ, Mikovits JA, Switzer WM, Ruscetti FW, Lo SC, Klimas N, Komaroff AL, Montoya JG, Bateman L, Levine S, Peterson D, Levin B, Hanson MR, Genfi A, Bhat M, Zheng H, Wang R, Li B, Hung GC, Lee LL, Sameroff S, Heneine W, Coffin J, Hornig M, Lipkin WI.** 2012. A multicenter blinded analysis indicates no association between chronic fatigue syndrome/myalgic encephalomyelitis and either xenotropic murine leukemia virus-related virus or polytropic murine leukemia virus. mBio **3:**e00266–12. doi:10.1128/mBio.00266-12.

23. **Paprotka T, Delviks-Frankenberry KA, Cingoz O, Martinez A, Kung HJ, Tepper CG, Hu WS, Fivash MJ, Jr, Coffin JM, Pathak VK.** 2011. Recombinant origin of the retrovirus XMRV. Science **333:**97–101.

24. **Hue S, Gray ER, Gall A, Katzourakis A, Tan CP, Houldcroft CJ, McLaren S, Pillay D, Futreal A, Garson JA, Pybus OG, Kellam P, Towers GJ.** 2010. Disease-associated XMRV sequences are consistent with laboratory contamination. Retrovirology **7:**111.

25. **Oakes B, Tai AK, Cingoz O, Henefield MH, Levine S, Coffin JM, Huber BT.** 2010. Contamination of human DNA samples with mouse DNA can lead to false detection of XMRV-like sequences. Retrovirology **7:**109.

26. **Sato E, Furuta RA, Miyazawa T.** 2010. An endogenous murine leukemia viral genome contaminant in a commercial RT-PCR kit is amplified using standard primers for XMRV. Retrovirology **7:**110.

27. **Fields BN, Knipe DM, Howley PM** (**ed**). 2013. Fields virology, 6th ed Wolters Kluwer/Lippincott Williams & Wilkins Health, Philadelphia, PA.

28. **Li L, Kapoor A, Slikas B, Bamidele OS, Wang C, Shaukat S, Masroor MA, Wilson ML, Ndjango JB, Peeters M, Gross-Camp ND, Muller MN, Hahn BH, Wolfe ND, Triki H, Bartkus J, Zaidi SZ, Delwart E.** 2010. Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. J. Virol. **84:**1674–1682.

29. **Rosario K, Marinov M, Stainton D, Kraberger S, Wiltshire EJ, Collings DA, Walters M, Martin DP, Breitbart M, Varsani A.** 2011. Dragonfly cyclovirus, a novel single-stranded DNA virus discovered in dragonflies (Odonata: Anisoptera). J. Gen. Virol. **92:**1302–1308.

30. **Hatakka A, Klein J, He R, Piper J, Tam E, Walkty A.** 2011. Acute hepatitis as a manifestation of parvovirus B19 infection. J. Clin. Microbiol. **49:**3422–3424.

31. **Mogensen TH, Jensen JM, Hamilton-Dutoit S, Larsen CS.** 2010. Chronic hepatitis caused by persistent parvovirus B19 infection. BMC Infect. Dis. **10:**246.

32. **Resendes AR, Majo N, van den Ingh TS, Mateu E, Domingo M, Calsamiglia M, Segales J.** 2011. Apoptosis in postweaning multisystemic wasting syndrome (PMWS) hepatitis in pigs naturally infected with porcine circovirus type 2 (PCV2). Vet. J. **189:**72–76.

33. **Xu B, Zhi N, Hu G, Wan Z, Zheng X, Liu X, Wong S, Kajigaya S, Zhao K, Mao Q, Young NS.** 2013. Hybrid DNA virus in Chinese patients with seronegative hepatitis discovered by deep sequencing. Proc. Natl. Acad. Sci. U. S. A. **110:**10264–10269.

34. **Mosley JW, Aach RD, Hollinger FB, Stevens CE, Barbosa LH, Nemo GJ, Holland PV, Bancroft WH, Zimmerman HJ, Kuo G, et al.** 1990. Non-A, non-B hepatitis and antibody to hepatitis C virus. JAMA **263:**77–78.

35. **Glaser CA, Honarmand S, Anderson LJ, Schnurr DP, Forghani B, Cossen CK, Schuster FL, Christie LJ, Tureen JH.** 2006. Beyond viruses: clinical profiles and etiologies associated with encephalitis. Clin. Infect. Dis. **43:**1565–1577.

36. **Swei A, Russell BJ, Naccache SN, Kabre B, Veeraraghavan N, Pilgard MA, Johnson BJ, Chiu CY.** 2013. The genome sequence of lone star virus, a highly divergent bunyavirus found in the Amblyomma americanum Tick. PLoS One **8:**e62083. doi:10.1371/journal.pone.0062083.

37. **Victoria JG, Kapoor A, Li L, Blinkova O, Slikas B, Wang C, Naeem A, Zaidi S, Delwart E.** 2009. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J. Virol. **83:**4642–4651.

38. **Li L, Pesavento PA, Leutenegger CM, Estrada M, Coffey LL, Naccache SN, Samayoa E, Chiu C, Qiu J, Wang C, Deng X, Delwart E.** 2013. A novel bocavirus in canine liver. Virol. J. **10:**54.

39. **Michael NL, Herman SA, Kwok S, Dreyer K, Wang J, Christopherson C, Spadoro JP, Young KK, Polonis V, McCutchan FE, Carr J, Mascola JR, Jagodzinski LL, Robb ML.** 1999. Development of calibrated viral load standards for group M subtypes of human immunodeficiency virus type 1

and performance of an improved Amplicor HIV-1 Monitor test with isolates of diverse subtypes. J. Clin. Microbiol. **37:**2557–2563.

40. **Yu G, Greninger AL, Isa P, Phan TG, Martinez MA, de la Luz Sanchez M, Contreras JF, Santos-Preciado JI, Parsonnet J, Miller S, DeRisi JL, Delwart E, Arias CF, Chiu CY.** 2012. Discovery of a novel polyomavirus in acute diarrheal samples from children. PLoS One **7:**e49449. doi:10.1371/journal.pone.0049449.

41. **Herman EK, Greninger AL, Visvesvara GS, Marciano-Cabral F, Dacks JB, Chiu CY.** 2013. The mitochondrial genome and a 60-kb nuclear DNA segment from Naegleria fowleri, the causative agent of primary amoebic meningoencephalitis. J. Eukaryot. Microbiol. **60:**179–191.

42. **Chiu CY, Yagi S, Lu X, Yu G, Chen EC, Liu M, Dick Jr, Carey EJKD, Erdman DD, Leland MM, Patterson JL.** 2013. A novel adenovirus species associated with an acute respiratory outbreak in a baboon colony and evidence of coincident human infection. mBio **4:**e00084–13. doi:10.1128/mBio.00084-13.

43. **Ruby JG, Bellare P, Derisi JL.** 2013. PRICE: software for the targeted assembly of components of (meta)genomic sequence data. G3 (Bethesda) **3:**865–880.

44. **Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A.** 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28:**1647–1649.

45. **Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW.** 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience **1:**18.

46. **Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J, Wooley J.** 2011. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. Nucleic Acids Res. **39:**D546–D551.

47. **Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA.** 2008. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics **9:**386.

48. **Guindon S, Delsuc F, Dufayard JF, Gascuel O.** 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol. Biol. **537:**113–137.

49. **Anisimova M, Gascuel O.** 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. **55:**539–552.

50. **Lamberg-Allardt C, Wilska M, Saraste KL, Gronlund T.** 1990. Vitamin D status of ambulatory and nonambulatory mentally retarded children with and without carbamazepine treatment. Ann. Nutr. Metab. **34:**216–220.

51. **Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH.** 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. **39:**D225–D229.

52. **Zadori Z, Szelei J, Lacoste MC, Li Y, Gariepy S, Raymond P, Allaire M, Nabi IR, Tijssen P.** 2001. A viral phospholipase A2 is required for parvovirus infectivity. Dev. Cell **1:**291–302.

53. **Shan T, Li L, Simmonds P, Wang C, Moeser A, Delwart E.** 2011. The fecal virome of pigs on a high-density farm. J. Virol. **85:**11697–11708.

54. **Monier A, Welsh RM, Gentemann C, Weinstock G, Sodergren E, Armbrust EV, Eisen JA, Worden AZ.** 2012. Phosphate transporters in marine phytoplankton and their viruses: cross-domain commonalities in viral-host gene exchanges. Environ. Microbiol. **14:**162–176.

55. **Kapoor A, Hornig M, Asokan A, Williams B, Henriquez JA, Lipkin WI.** 2011. Bocavirus episome in infected human tissue contains non-identical termini. PLoS One **6:**e21362. doi:10.1371/journal.pone.0021362.

56. **Chen CL, Jensen RL, Schnepp BC, Connell MJ, Shell R, Sferra TJ, Bartlett JS, Clark KR, Johnson PR.** 2005. Molecular characterization of adeno-associated viruses infecting children. J. Virol. **79:**14781–14792.

57. **Boom R, Sol CJ, Salimans MM, Jansen CL, Wertheim-van Dillen PM, van der Noordaa J.** 1990. Rapid and simple method for purification of nucleic acids. J. Clin. Microbiol. **28:**495–503.

58. **Erlwein O, Robinson MJ, Dustan S, Weber J, Kaye S, McClure MO.** 2011. DNA extraction columns contaminated with murine sequences. PLoS One **6:**e23484. doi:10.1371/journal.pone.0023484.

59. **Evans GE, Murdoch DR, Anderson TP, Potter HC, George PM, Chambers ST.** 2003. Contamination of Qiagen DNA extraction kits with Legionella DNA. J. Clin. Microbiol. **41:**3452–3453.

60. **Qiu X, Swanson P, Tang N, Leckie GW, Devare SG, Schochetman G, Hackett J, Jr.** 2012. Seroprevalence of xenotropic murine leukemia virus-related virus in normal and retrovirus-infected blood donors. Transfusion **52:**307–316.

61. **Suttle CA.** 2005. Viruses in the sea. Nature **437:**356–361.

62. **Rosario K, Duffy S, Breitbart M.** 2009. Diverse circovirus-like genome architectures revealed by environmental metagenomics. J. Gen. Virol. **90:**2418–2424.

63. **Roekring S, Nielsen L, Owens L, Pattanakitsakul SN, Malasit P, Flegel TW.** 2002. Comparison of penaeid shrimp and insect parvoviruses suggests that viral transfers may occur between two distantly related arthropod groups. Virus Res. **87:**79–87.

64. **Amin SA, Parker MS, Armbrust EV.** 2012. Interactions between diatoms and bacteria. Microbiol. Mol. Biol. Rev. **76:**667–684.

65. **Nagasaki K.** 2008. Dinoflagellates, diatoms, and their viruses. J. Microbiol. **46:**235–243.

66. **Karp A, Ingram D, Isaac P.** 2001. Molecular tools for screening biodiversity plants and animals, vol pp xxiv. Kluwer Academic Publishers, Dordrecht, the Netherlands.

67. **Parsons RJ, Breitbart M, Lomas MW, Carlson CA.** 2012. Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. ISME J. **6:**273–284.

68. **Stramer SL, Hollinger FB, Katz LM, Kleinman S, Metzel PS, Gregory KR, Dodd RY.** 2009. Emerging infectious disease agents and their potential threat to transfusion safety. Transfusion **49**(Suppl 2)**:**1S–29S.