Title

Non-invasive surveys of mammalian viruses using environmental DNA

Highlights

- Environmental DNA (water and blood-sucking leeches) provided a non-invasive method of screening wildlife for viruses
- A comprehensive viral RNA oligonucleotide bait set was developed to capture known and unknown mammalian virus diversity
- Leech blood meal host determination and viruses identified were congruent
- Viruses determined from water correlated with known and observed species visiting the water sources

In brief

Alfano, Dayaram, et al. demonstrate that environmental DNA from southeast Asian leech bloodmeals and waterholes from Africa and Mongolia can be used as to detect viruses circulating in wildlife. These nucleic acid sources may represent an effective non-invasive resource for studying wildlife viral diversity and emerging viruses pre-emergence.

Non-invasive surveys of mammalian viruses using environmental DNA

Niccolo Alfano[1,2]*, Anisha Dayaram[3,4]*, Jan Axtner[1], Kyriakos Tsangaras[5], Marie-Louise Kampmann[1,6], Azlan Mohamed[1,7], Seth T. Wong[1], M. Thomas P. Gilbert[8,9] Andreas Wilting[1,11]**, Alex D. Greenwood[3,10,11]**

Affiliations

[1] Department of Ecological Dynamics, Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Str. 17, 10315 Berlin, Germany

[2] Department of Biodiversity & Molecular Ecology, Fondazione Edmund Mach, Research and Innovation Centre, Via Edmund Mach 1, 38010 San Michele All'adige (TN), Italy

[3] Department of Wildlife Diseases, Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Str. 17, 10315 Berlin, Germany

[4] Charité-Universitätsmedizin Berlin, corporate member of Freie Universitäts Berlin and Humboldt-Universität of Berlin, Institut für Neurophysiologie, Berlin, Germany

[5] Department of Translational Genetics, The Cyprus Institute of Neurology and Genetics, Iroon Avenue 6, Agios Dometios 2371, Cyprus

[6] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

[7] WWF-Malaysia, PJCC, 46150 Petaling Jaya, Selangor, Malaysia.

[8] The GLOBE Institute, University of Copenhagen, Øster Farimagsgade 5A, 1352 Copenhagen, Denmark

[9] University Museum. NTNU, 7491 Trondheim, Norway

[10] Department of Veterinary Medicine, Freie Universität Berlin, Robert-von-Ostertag-Str. 7-13, Berlin 14163, Germany

[11] Lead contacts

* These authors contributed equally

** Correspondence: wilting@izw-berlin.de (**AW**), greenwood@izw-berlin.de (**ADG**)

Summary

Environmental DNA (eDNA) and its subdiscipline, invertebrate-derived DNA (iDNA) have been used to survey biodiversity non-invasively [1,2]. Water is ubiquitous in most ecosystems, and, among invertebrates, terrestrial haematophagous leeches are abundant and can be easily collected in many tropical rainforests [3,4]. Such non-invasive nucleic acid sources can mitigate difficulties of obtaining wildlife samples, particularly in remote areas or for rare species. Recently, eDNA/iDNA sources have been applied to monitoring specific wildlife pathogens [5,6]. However, previous studies have focused on known pathogens, whereas most wildlife pathogens are uncharacterized and unknown. Non-invasive approaches to monitoring known and novel pathogens may be of particular benefit in ecosystems prone to viral emergence, many of which occur in areas where invasive sampling is challenging, such as tropical rainforests. Here, we show that both eDNA from natural waterholes, and iDNA from terrestrial haematophagous leeches, can be used to detect unknown viruses circulating in mammalian hosts (Figure 1). Using a curated set of RNA oligonucleotides based on the ViroChip microarray assay [7] as baits in a hybridization capture system, multiple mammalian RNA and DNA viruses were detected from both eDNA and iDNA samples. Congruence was found between host DNA assignment and viruses identified in leeches, and between animals observed visiting the waterholes and the viruses detected. Our results demonstrate that eDNA/iDNA samples may represent an effective non-invasive resource for studying wildlife viral diversity. Several of the detected viruses were novel, highlighting the potential of eDNA/iDNA for epidemiological analysis of emerging viruses prior to their emergence.

Results

Positive and negative controls

In order to test the sensitivity of the viral capture in recovering vertebrate host viruses, the capture system was first applied to a positive control consisting of medical leeches fed with human blood spiked with two RNA viruses and two DNA viruses at different concentrations [8]. All four viruses were detected, even if enrichment efficiency (proportion of on-target viral reads) and target genome recovery varied among viruses (Suppl. Fig. 1). No viral contigs were identified in the negative controls included to monitor laboratory contaminations for either the leech or water experiments.

Leech viral identification

Tiger leeches (*Haemadipsa picta*) and brown leeches (*Haemadipsa zeylanica*) were collected in Malaysian Borneo and processed as pools (bulk samples) consisting of 1 to 77 individual leeches separated by leech species and sampling location. Viruses were identified in 40 of the 68 leech pools analysed (59%) (Fig. 2; Suppl. Tab. 1). In 18 of these (45%), two to three viruses were identified. Sequence data from six vertebrate-infecting viral families were detected, including the *Anelloviridae*, *Circoviridae*, *Coronaviridae*, *Parvoviridae*, *Retroviridae* and *Rhabdoviridae*. The most common viral group detected was *Rhabdoviridae* which was found in 37% of samples (25 samples out of 68), followed by *Coronaviridae* which was identified in 24% of samples (16 samples). Members of the *Anelloviridae* were identified in 12% of samples (8 samples), *Retroviridae* in three samples (4%), and *Parvoviridae* and *Circoviridae* in two samples (3%) (Fig. 2; Suppl. Tab. 1).

Rhabdoviridae contigs were genetically similar to three different viral genera (Suppl. Tab. 1). Five contigs were most similar (69-77%) to the Vesicular stomatitis Indiana virus (VSIV) (genus *Vesiculovirus*) as determined by BLAST searches. The limited similarity of the contigs to known rhabdoviruses suggest they may represent a new genus related to fish rhabodviruses (*Perhabdovirus* and *Sprivirus*) or *Vesiculovirus* (Suppl. Fig. 2). The other contigs clustered phylogenetically, suggesting they represent two new species of a rhabdovirus related to lyssaviruses (Suppl. Fig. 2). Although in most cases one contig per sample was observed, in five samples (L4, L12, L23, L58, L68) two different viruses were found. Several viral regions for *Rhabdoviridae* were represented in the baits. However, most of the oligonucleotides were specific for the L gene which encodes the RNA-dependent RNA polymerase. All the recovered contigs mapped to the L gene (Suppl. Fig. 3A-C). The viral contig sequences were confirmed by PCR and Sanger sequencing for L55 and L58 (Suppl. Fig. 3D).

All *Coronaviridae* contigs matched a bat betacoronavirus as determined by BLAST searches with identities between 70-73% (Suppl. Tab. 1). The resulting sequence did not cluster in any of the four clades representing the known *Coronaviridae* genera, suggesting it may represent a novel coronavirus genus (Suppl. Fig. 4). Each contig overlapped with the coronavirus *RNA-dependent RNA polymerase* gene (orf1ab), the viral region mainly targeted by the RNA oligonucleotide baits (Suppl. Fig. 5).

Anelloviridae contigs matched either porcine torque teno virus (PTTV) (95-96% identity), a giant panda anellovirus (GpAV) (81-92% identity) or a masked palm civet torque teno virus (Pl-TTV) (83-92% identity) (Suppl. Tab. 1). The PTTV contigs were found in two samples (L8 and L37), while the GpAV and Pl-TTV contigs were detected in six samples. GpAV was the best match in four samples (L7, L17, L36, L39) and Pl-TTV in three (L21, L25, L39). In sample L39 both were identified. Every *Anelloviridae* contig mapped to the non-coding region of the relative reference genome since all *Anelloviridae* baits targeted the same untranslated region (Suppl. Fig. 6A, C, E). The non-coding region sequenced is not phylogenetically informative and therefore, phylogenetic analysis could not be performed. Viral contigs were confirmed by PCR and Sanger sequencing for samples L7, L17, L25 and L37 (Suppl. Fig. 6B, D, F).

Three *Circoviridae* contigs matching a porcine circovirus (PCV) (100% identity) were identified in L7 and L59 (Fig. 1; Suppl. Tab. 1). Two non-overlapping but adjacent contigs were retrieved from L7. A single contig overlapping with one of the two contigs determined from L7 was recovered from L59 (Suppl. Fig. 7A). The contigs mapped to the PCV replication protein (Rep), targeted by the *Circoviridae* baits (Suppl. Fig. 7A). The two overlapping contigs of L7 and L59 were confirmed by PCR and Sanger sequencing (Suppl. Fig. 7B). Since the identity of the contigs with known viral sequences in GenBank was 100%, no phylogenetic analysis was performed.

Parvoviridae contigs with the highest similarity to porcine parvovirus (PPV) were found in L8 (1 contig with 98% identity) and L14 (2 contigs with 74-77% identity) (Suppl. Tab. 1). The contig of L8 clustered within the *Tetraparvovirus* genus, close to ungulate parvoviruses (porcine, ovine and bovine PV), while the contigs of L14 within the *Copiparvovirus* genus, close to PPV4 (Suppl. Fig. 8). Two of the three contigs mapped to the *replicase* gene, while one from L14 mapped to an intergenic region (Suppl. Fig. 7C). Whereas the *replicase* region of PPV was covered by *Parvoviridae* baits, the intergenic region was not (Suppl. Fig. 7C). This portion of the virus may have been recovered by other non-*Parvoviridae* baits targeting that region non-specifically.

Retroviridae contigs similar to the simian and feline foamy virus (*Spumaretrovirinae* subfamily, 79-82% identity) were detected in three samples (L7, L46, L64) (Suppl. Tab. 1). Phylogenetically the contigs clustered together as a sister group to the feline foamy viruses (*Felispumavirus* genus), potentially being a new genus within the *Spumaretrovirinae* (Suppl. Fig. 9). The contigs mapped to the *polymerase* gene, which the exogenous retrovirus baits were designed to target (Suppl. Fig. 7D).

Leech bloodmeal host assignments

The mammalian hosts of the leeches were determined by metabarcoding [31]. The bearded pig (*Sus barbatus*) was identified in samples yielding porcine viruses, such as porcine circovirus (L7), porcine parvovirus (L8) and porcine torque teno virus (L8 and L37). Four leech samples with giant panda anellovirus (L7, L17, L36, L39) sequences yielded sun bear (*Helarctos malayanus*) sequences. Malay civet (*Viverra tangalunga*) was identified in one of the three samples (L25) with masked palm civet torque teno virus sequences. Fourteen of the 16 samples with the potentially new coronavirus genus (87.5%) yielded deer sequences, specifically sambar (*Rusa unicolor*), indicating that the novel coronavirus might be a cervid virus. Similarly, the novel *Lyssavirus*-like *Rhabdoviridae* sequences were associated with cervid species (sambar or muntjac) (16 of 22 samples, 73%). However, due to the high prevalence of deer in the samples tested (70%) we could not reject that the occurrence of viruses and deer are independent variables ($Chi^2_{Coronaviridae} = 1.916$, 1 df, p = 0.1663; $Chi^2_{Rhabdoviridae} = 1.046$, 1 df, p = 0.3064).

Waterhole viral identification

Five waterholes from Tanzania and six from Mongolia were tested. From each waterhole, one water filtrate and one sediment sample were collected (except for one waterhole where only a sediment sample was collected), for a total of twenty-one samples. Five samples (two water and three sediment samples) in total were positive for viral sequences (23.8%). Four viral families were identified including: *Retroviridae, Herpesviridae, Adenoviridae* and *Papillomaviridae*. In filtered water and sediment samples collected from the same waterhole, only one virus per sample was generally identified and in one location (WM20 and SM20) contigs from different viral families were isolated based on sample type. Differences between sediment and water are not unexpected as the sediment likely represents a longer-term accumulation of biomaterial and the water represents more acute contamination at the surface and variable mixing throughout.

Of the 11 water filtrate samples tested, two samples from Mongolia (WM3 and WM20) (18.2%) had viral contigs with 100% identity to the Equid herpesvirus 1 and 3 (EHV-1 and EHV-3). The contig of WM20 mapped to the membrane glycoprotein B, whereas the two contigs of WM3 to the DNA packaging protein and membrane glycoprotein G, all regions covered by the *Herpesviridae* baits (Suppl. Fig. 10A-D). A nested panherpes PCR targeting the *DNA polymerase* gene and the resulting Sanger sequences further confirmed EHV presence (Suppl. Fig. 10E). Several equine species including domestic horses inhabit the Gobi Desert [9], which is consistent with the presence of these viruses.

From the 12 sediment samples tested, two from Mongolia and one from Tanzania yielded viral sequences (25%) representing three viral families including: *Retroviridae, Adenoviridae* and *Papillomaviridae.* Mongolian sediment sample SM6 was positive for four contigs mapping to the *protease* (*pro*) gene of the Jaagsiekte sheep retrovirus (JSRV) with 100% identity (Suppl. Fig. 11A). JSRV from this sample was further confirmed by PCR (Suppl. Fig. 11A). Mongolia sediment sample SM20 was positive for Equine adenovirus (100% identity) with a contig mapping to a region comprised between the *pVI* and *hexon* capsid genes (Suppl. Fig. 11B). Given that multiple equine species are found in the Gobi Desert in Mongolia, it is likely that the water sources sampled may have been frequented by these species [10]. The sediment sample from Tanzania ST38 was positive for a Zetapapillomavirus related to the *Equus caballus* papillomavirus and *Equus asinus* papillomavirus (74% identity; *E1-E2* genes) (Suppl. Fig. 11C; Suppl. Fig. 12), consistent with the detection of Plains zebra's (*Equus quagga*) DNA from this water source [3]. Given that both captive and wild zebras have been known to contract bovine papillomaviruses [10,11] it is likely that they are susceptible to different equine papillomaviruses.

Discussion

Emerging infectious viruses increasingly threaten human, domestic animal and wildlife health [12]. Sixty percent of emerging infectious diseases in humans are of zoonotic origin [13]. Wildlife trade and consumption of bushmeat, especially in Africa and Asia, have increasingly played a role in pathogen spillovers into human populations [14,15]. Wildlife markets have recently facilitated the spillover of SARS-CoV-2 to humans [16] resulting in a pandemic [17]. The 2002–2003 SARS-CoV outbreak [18], the Ebola outbreak in West Africa [19] and the global emergence of HIV [20] have all been linked to wildlife trade and bushmeat consumption. Early detection of novel infectious agents in wildlife represent a key factor to prevent their emergence. However, identification, surveillance and monitoring of emerging viruses using currently broadly applied approaches based on direct sampling of wildlife requires enormous investment in sampling, particularly for viruses that have low prevalence [21]. For example, 25,000 wild birds were sampled in Germany to detect avian influenza prevalence below 1% [22]. Similarly, sampling of over 8,157 animals in Poland was required to determine an 0.12% prevalence of African swine fever virus (ASF) [23]. Sampling under remote field conditions or in developing countries present additional challenges.

We provide evidence that environmental and invertebrate-derived DNA samples including waterhole water, sediment and wild haematophagous terrestrial leeches can be used to survey known and unknown viruses. DNA and RNA viruses could be detected in 59% and 20.8% of the iDNA (leech) and eDNA (waterhole) samples, respectively. The congruence of host DNA assignment for leeches and viral families identified suggests that bloodmeals are a useful resource for determining viral diversity. Similarly, the detection of primarily equine viruses from African and Mongolian waterholes, where intense wild equid visitation rates were directly observed, suggests eDNA viruses from this resource reflect host utilization of the water and do not derive from other environmental sources such as fomites distributed over long distances.

PCR based approaches, as used in earlier studies to detect pathogens from flies [5,24] or, under laboratory conditions, in medicinal leeches [8], require prior knowledge about the expected pathogens in the samples. The unknown viral diversity in the wild, and the potential degradation of viral nucleic acids in bloodmeals or in the environment, may affect detection by PCR resulting in high false negative rates. RNA oligonucleotide based hybridization capture overcomes such limitations because the short baits can capture divergent and degraded DNA. The comprehensive viral group representation in the RNA bait set also allows for the determination of both viral presence and viral diversity with a relatively simple workflow. The ability of oligonucleotides with substantial divergence from the target sequence to capture more distantly related sequences is particularly useful in virology since most viruses are uncharacterized in wildlife and many evolve rapidly.

Using short RNA baits to capture highly conserved sequences from every known vertebrate viral genome is a useful and relatively inexpensive approach for providing an initial viral identification. However, to fully characterize each virus, the RNA oligonucleotide bait set would need to be customized to retrieve full length viral genomes. Initial screening with full length genomes for all viruses is costly and may result in detection of host DNA in cases of spurious homology between viruses and host DNA sequence.

Several novel viruses were identified with our short RNA bait approach, which is not unexpected as little is known about the virology of wildlife in Southeast Asia, where the leeches were collected. Several viral contigs were phylogenetically distinct from known viruses and may represent new genera. For example, the novel coronavirus identified in leech bloodmeals did not cluster with any of the known *Coronaviridae* clades. This finding highlights the ability of this method to detect unknown viruses. We could also associate the novel corona- and rhabdoviruses with mammal bloodmeals with limited evidence of a cervid association for both. Cervids are regularly sold as bushmeat in wildlife markets [25] and both recent coronavirus epidemics (SARS-CoV [18] and SARS-Cov-2 [16]) spilled over from wildlife. This suggests that eDNA/iDNA-based pathogen surveillance approaches may complement efforts to proactively identify viruses that could potentially spillover to humans or livestock.

The collection of wild haematophagous invertebrates such as leeches or water and sediments has both advantages and disadvantages compared to invasively collected wildlife samples. Large amounts of DNA can be extracted from bloodmeals, in particular when leeches are processed in bulk. We pooled up to 77 leeches and many of our leech bulk samples contained a diverse mix of mammalian DNA. A disadvantage of leeches is that they cannot be found in all environments: for example haematophagous terrestrial species are restricted to tropical rainforests of Asia, Madagascar and Australia [26]. In addition, leech feeding biases could influence diversity surveys [4,27]. However, this disadvantage could be overcome in the future by employing additional invertebrates such mosquitoes [28] or carrion flies [29]. Waterholes are commonly found in almost all environments. In environments with seasonal water shortages, DNA from animals can become highly concentrated due to many animals utilizing rare water sources. The disadvantages are that the dilution factor of water, depending on water body size, can obscure rare DNA sequences and mixed host species sequences are generally the rule rather than the exception. Further experiments with field filtration and sample concentration such as methods used with pathogen detection in waste water may improve detection rates [30].

Environmental DNA and in particular its subdiscipline invertebrate-derived DNA viral hybridization capture may be a useful and economical tool for identifying and characterizing major viral pathogens particularly in difficult to access sampling environments prior to viral emergence. Sampling in environments where direct access to animals is difficult or highly restricted, eDNA and iDNA may be the only option to detect viral pathogens in the wild. The current study suggests this approach will be successful in either complementing or replacing invasive approaches.

Star methods

Sample collection

Leeches

Two types of leeches, tiger leeches (*Haemadipsa picta*) and brown leeches (*Haemadipsa zeylanica*) were collected from February to May 2015 in the Deramakot Forest Reserve in Sabah, Malaysian Borneo as described in Abrams et al. 2019 [27] and Axtner et al. 2019 [31]. All leeches of the same type (tiger or brown) from the same site and occasion were pooled and processed as one sample. Number of leeches ranged from 1 to 77 per pool (median= 7). Samples were stored in RNA fixating saturated ammonium sulfate solution and exported under the permit ‘JKM/MBS.1000-2/3 JLD.2 (8)’ issued by the Sabah Biodiversity Council. A total of 68 pools (L1-L68) were selected for viral capture to maximize representation of host wildlife species identified from bloodmeals [31].

Sediment and water

In February, June, July and October 2016 samples were collected from the Serengeti National Park (ca. 2.2° S, 34.8° E) Tanzania from waterholes. In October 2015 samples were collected from South East Gobi (45.5905°N, 107.1596 °E) and in between June – July 2016 samples were collected from

Gobi B (45.1882°N, 93.4288°E) in Mongolia. At each waterhole, 50 ml of water was passed through a 0.22 µm Sterivex filter unit (Millipore) using a disposable 50-ml syringe to remove debris from water. In addition, 25 g of the top 1-3 cm of sediment was collected at each waterhole. The samples were stored on ice packs during the respective field trip, and frozen at -20 °C. In total water filtrate and sediment samples were sampled at 12 waterholes, six respectively from Mongolia and Tanzania. For each sample, 32 ml of water filtrate was ultra-centrifuged at 28,000 rpm for 2 hours to pellet DNA and viral particles. The supernatant was then removed, the pellet re-suspended in 1 ml of cold phosphate-buffered saline (PBS) (pH 7.2) (Sigma-Aldrich) and left at 4 °C overnight.

Preparation of samples and nucleic acid extraction

Leeches

Leeches were cut into small pieces with a new scalpel blade and lysed overnight (≥12 hours) at 55°C in proteinase K and ATL buffer at a ratio of 1:10; 0.2 ml per leech. Total nucleic acids were extracted from leech samples using the DNeasy 96 Blood & Tissue kit (Qiagen) following the manufacturer's protocol and finally eluted twice with 100µl 1x Tris-EDTA buffer. For more detailed information on laboratory protocol and samples please see Axtner et al. 2019 and its supporting data (doi: 10.5524/100570).

Water and sediment

500 µl of the centrifuged filtrate was used to extract viral nucleic acids using the RTP$^{®}$ DNA/RNA Virus Mini Kit (Stratec biomedical). The following modifications were made to the original protocol: 400 µl of lysis buffer, 400 µl of binding buffer and 20 µl of proteinase K and carrier RNA were used per sample. Samples were eluted in 60 µl. The NucleoSpin Soil kit (Macherey-Nagel) was used to extract DNA/RNA from sediment. 500 mg of soil was extracted according to the manufactures protocol using an elution volume of 100 µl.

Positive control

As a positive control medical leeches (*Hirudo* spp.) were fed human blood spiked with four viruses [8]. Two RNA viruses, influenza A and measles morbillivirus, and two DNA viruses, bovine herpesvirus and human adenovirus were used (see Kampmann et al. 2017 [8] for details).

Library Preparation

The RNA was reverse transcribed using SuperScript III and IV (Thermo Fisher Scientific) with random hexamers prior to second-strand synthesis with Klenow fragment (New England Biolabs). The resulting double-stranded cDNA/DNA mix was sheared to an average fragment size of 200 bp using a M220 focused ultrasonicator (Covaris). Sheared product was purified using the ZR-96 DNA Clean & Concentrator-5 kit (Zymo). Dual-indexed Illumina sequencing libraries were constructed as described by Meyer and Kircher 2010 [32] with the modifications described in Alfano et al. 2015 [33]. Each library was amplified in three replicate reactions to minimize amplification bias in individual PCRs. The three replicate PCR products for each sample were pooled and purified using the MinElute PCR Purification Kit (Qiagen). Negative control libraries were also prepared from different stages of the experimental process (extraction, reverse transcription, library preparation and index PCR) and indexed separately to monitor any contamination introduced during the experiment. Amplified libraries were quantified using the 2200 TapeStation (Agilent Technologies) on D1000 ScreenTapes.

RNA oligonucleotide Bait Design

The targeted sequence capture panel was designed based on the oligonucleotide probes represented on the Virochip microarray [34]. The Virochip is a pan-viral DNA microarray comprising the most highly conserved 70 mer sequences from every fully sequenced reference viral genome present in GenBank, which was developed for the rapid identification and characterization of novel viruses and emerging infectious disease. We retrieved the viral oligonucleotides from the 5th generation Virochip (Viro5) [35], which are publicly available at NCBI's Gene Expression Omnibus (GEO) repository [36], accession number GPL13323 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL13323). This platform includes ~17,500 oligonucleotides (70 mer nucleotides) derived from ~2,000 viral species. We excluded sequences from bacteriophage, plant viruses, viral families infecting only invertebrates and endogenous retroviruses. We included viruses that could have both vertebrate and invertebrate hosts, such as vertebrate viruses with insect vectors. Exogenous retroviruses were represented but murine leukemia viruses (MLVs) were removed. MLVs sequences may interfere with the capture of other viruses, since MLVs can cross enrich endogenous retroviruses which can represent large portions of several vertebrate genomes and mask rarer viral sequences. Control oligonucleotides included in the Virochip, such as those from human genes, yeast intergenic sequences, and human papilloma virus sequences present in HeLa cells were also removed. Ninety-two 70-mer oligonucleotides covering (spaced end-to-end) the entire *pol* and *gB* genes of Equine herpesvirus 1 (EHV-1) were included as PCR screening of the water samples indicated they were positive for this virus (data not shown). The resulting 13,532 oligonucleotides were examined for repetitive elements, short repeats, and low complexity regions, which are problematic for probe design and capture, using RepeatMasker. Repetitive motifs were identified in 234 oligonucleotides, which were removed. The final targeted sequence capture panel consisted of 13,298 unique sequences which were synthesized (as a panel of biotinylated RNAs) at MYcroarray (Ann Arbor, USA).

Viral Enrichment Strategy and Sequencing

In-solution target enrichment via hybridization-based capture was performed according to the manufacturer's protocol (MYbaits® custom targeted enrichment, MYcroarray, Ann Arbor, USA), with the following modifications for likely partially degraded samples with an expected low target viral content: 50uL Dynabeads® M-270 Streptavidin beads (Invitrogen) instead of 30 uL Dynabeads® MyOne™ Streptavidin C1 (Invitrogen); hybridization, bead-bait binding, and wash steps temperature set to 60°C; 48 hours hybridization time; 200 ng baits per reaction; 10 µL indexed library inputs. For capture, libraries generated from pooled leeches consisting of more than 16 individuals were captured individually, while libraries generated from pools of fewer individuals were combined to have a comparable number (15-20) of leeches per capture. This was done in order to ensure that each individual leech represented in each library was allocated enough bait for capture. For capture libraries generated from water and sediment samples. Samples were pooled in groups of two. Sediment and water cDNA and DNA were pooled separately. Per pooled sample, 5 µl of baits were used to ensure enough bait for each sample. The enriched libraries were re-amplified using Herculase II Fusion DNA polymerase (Agilent Technologies) with P5 and P7 Illumina library outer primers with the same cycling conditions described in Alfano et al. 2016. The re-amplified enriched libraries were purified using the MinElute PCR Purification Kit (Qiagen), quantified using the 2200 TapeStation (Agilent Technologies) on D1000 ScreenTapes and finally pooled in equimolar amounts for single-read sequencing on two lanes of an Illumina NextSeq 500 with the TG NextSeq® 500/550 High Output Kit v2 (300 cycles).

Data analysis and bioinformatics pipelines

A total of 219,580,903 sequence reads 300 bp long were generated (average: 3,181,781 single reads per sample; standard deviation [SD]: 1,481,098) (Suppl. Tab. 1) and sorted by their dual index sequences. Cutadapt v1.16 and Trimmomatic v0.36 were used to remove adapter sequences and low-quality reads using a quality cutoff of 20 and a minimal read length of 30 nt. After trimming, 97% of the sequences were retained. Three different approaches (A, B, C) were used to analyse the viral capture data:

A) Leech reads were removed from the dataset by alignment to the *Helobdella robusta* genome v1.0 (assembly GCA_000326865.1), which is the only complete genome of Hirudinea available in GenBank, and all leech sequences from GenBank (4,957 sequences resulting from "Hirudinea" search) using Bowtie2 v2.3.5.1 [37]. This left 81% of the original reads (Suppl. Tab. 1). Then, the filtered reads were searched by BLAST against a database generated from the capture bait sequences. The reads which matched with baits were then extracted and screened against the entire NCBI nucleotide database (nt) using BLASTn to find the best viral match. The filtered reads were mapped both to the corresponding bait sequence and the genome sequence of the best hit obtained by BLAST against the complete nt database, in order to generate a consensus sequence. This consensus sequence was again searched against the NCBI nt database using BLASTn to obtain a viral assignment.
B) Leech reads were removed as in method A. In addition, rRNA reads were removed using SortMeRNA [38], leaving 75% of the original reads (Suppl. Tab. 1). The filtered reads were *de novo* assembled using both Spades v3.11.1 [39] and Trinity v2.6.6 [40] assemblers. The obtained contigs from Spades and Trinity were pooled and clustered to remove duplicated or highly similar sequences using USEARCH v11.0.667 [41] with a 90% threshold identity value. The centroids were then subjected to sequential BLAST searches against the NCBI nucleotide database and NCBI RefSeq viral protein database using BLASTn and BLASTx, respectively.
C) The adaptor and quality trimmed data were uploaded to Genome Detective [42], a web base software that assembles viral genomes from NGS data. The software first groups reads into different buckets based on the proteins similarity to different viral hits. Genome detective then *de novo* assembles the reads of each bucket creating a longer consensus sequence that is then searched against the NCBI RefSeq viral database using BLASTx and BLASTn algorithms. The results of amino acid and nucleotide search are combined and viral hit is assigned based on the best combined score.

Bacteriophages, invertebrate viruses and retroviruses were excluded from subsequent steps, which only focused on eukaryotic, specifically vertebrate viruses. The results of the three methods were compared and the viral contigs obtained were manually inspected. If more than one method generated a contig with the same viral hit, the contigs from each method were compared. If they had the same sequence or were overlapping, the longest contig was selected. The filtered reads were mapped to the viral contigs to calculate the number of viral reads for each virus. Finally, the viral contigs were mapped to the reference genome of the virus corresponding to the best BLAST hit using Geneious v11.0.2 (Biomatters, Inc.) [43]. The baits were mapped to the same references to determine the genomic positions targeted by our bait panel for each virus.

Phylogenetic analyses

Viral contigs were assigned to viral families according to the best BLAST results. Comprehensive sets of representative sequences from these viral families were retrieved from GenBank and aligned with the contigs using MAFFT v7.450 [44]. Phylogenetic analysis was performed using the maximum-likelihood method based on the general time reversible substitution model with among-site rate heterogeneity modelled by the Γ distribution and estimation of proportion of invariable sites available in RAxML v8 [45], including 500 bootstrap replicates to determine node support. Phylogenetic analyses were performed only on viral contigs i) showing divergence from known viruses, i.e. with both BLAST identity and coverage to the best reference below 95%, to place them into a phylogenetic context, and ii) mapping to phylogenetically relevant genomic regions. Therefore, *Circoviridae* and *Anelloviridae* contigs were excluded as were those identified from water.

Leech vertebrate host assignments

Host identification of leeches followed an eDNA/iDNA workflow recently published [31]. In summary, leech samples were digested and short fragments of the mitochondrial markers 12S, 16S and cytochrome B were amplified in four PCR replicates each resulting in 12 PCR replicates per sample. We used a twin-tagging 2-step PCR protocol and PCR products were sequenced using an Illumina MiSeq (for details please see Axtner et al. 2019 [31]). After demultiplexing and read processing, each haplotype was taxonomically assigned to a curated reference database using PROTAX [46]. Taxonomic assignments followed the criteria of Axtner et al. 2019.

Viral detection confirmation by PCR

The primers listed in Suppl. Tab. 2 were designed to confirm by PCR the viral contig sequences generated by the three approaches (A, B, C see above) from the leech samples. For PCRs targeting RNA viruses, 50 uL of extract were digested with rDNase I (Ambion) following the manufacturer's protocol. The DNAse-digested extracts were then purified using the RNeasy MinElute Cleanup Kit (Qiagen). RNA was reverse transcribed into cDNA using iScript™ Reverse Transcription Supermix (Bio Rad). Sediment and water samples that tested positive for EHV and JSRV were screened using a previously described pan-herpes PCR [47] and for JSRV [48], respectively. The resulting amplicons were Sanger sequenced.

Acknowledgments

We would like to thank Joseph DeRisi for guidance in using his oligonucleotide data. This project received financial support from the German Federal Ministry of Education and Research (BMBF FKZ: 01LN1301A) to NA, JA, AM and AW and was supported by funds from the Leibniz Gemeinschaft, SAW-2015-IZW-1 440 to AD and ADG. We thank Peter Seeber and Sanatana-Erini Soilemetzidou for collection of water and sediment samples in Africa and Mongolia respectively. We thank the Sabah Forestry Department, especially Johnny Kissing, Peter Lagan, and Datuk Sam

Mannan, for supporting the fieldwork and the Sabah Biodiversity Council for providing research, collection, and export permits (JKM/MBS.1000-2/3 JLD.2) for the leech work.

Author Contributions

ADG and AW designed the study; AM and STW collected the leeches in the field; JA performed the leech nucleic acids extractions and the PCRs on leeches; NA and AD performed the capture experiments; NA and KT performed the bioinformatics and phylogenetic analyses; NA, AD, ADG and AW wrote the manuscript; NA, AD, JA, KT, MLK, TPG, AW and ADG reviewed the manuscript; all authors read and approved the final manuscript.

Declaration of Interests

The authors declare no competing interests.

References

1. Bohmann, K., Evans, A., Gilbert, M.T.P., Carvalho, G.R., Creer, S., Knapp, M., Douglas, W.Y., and De Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. Trends Ecol. Evol. *29*, 358–367.

2. Thomsen, P.F., and Willerslev, E. (2015). Environmental DNA–An emerging tool in conservation for monitoring past and present biodiversity. Biol. Conserv. *183,* 4–18.

3. Seeber, P.A., McEwen, G.K., Löber, U., Förster, D.W., East, M.L., Melzheimer, J., and Greenwood, A.D. (2019). Terrestrial mammal surveillance using hybridization capture of environmental DNA from African waterholes. Mol. Ecol. Resour.

4. Schnell, I.B., Sollmann, R., Calvignac-Spencer, S., Siddall, M.E., Douglas, W.Y., Wilting, A., and Gilbert, M.T.P. (2015). iDNA from terrestrial haematophagous leeches as a wildlife surveying and monitoring tool–prospects, pitfalls and avenues to be developed. Front. Zool. *12,* 24.

5. Gogarten, J.F., Düx, A., Mubemba, B., Pléh, K., Hoffmann, C., Mielke, A., Müller-Tiburtius, J., Sachse, A., Wittig, R.M., Calvignac-Spencer, S., *et al.* (2019). Tropical rainforest flies carrying pathogens form stable associations with social nonhuman primates. Mol. Ecol. *28,* 4242–4258.

6. Mosher, B.A., Huyvaert, K.P., Chestnut, T., Kerby, J.L., Madison, J.D., and Bailey, L.L. (2017). Design-and model-based recommendations for detecting and quantifying an amphibian pathogen in environmental samples. Ecol. Evol. *7*, 10952–10962.

7. Chen, E.C., Miller, S.A., DeRisi, J.L., and Chiu, C.Y. (2011). Using a Pan-Viral Microarray Assay (Virochip) to Screen Clinical Samples for Viral Pathogens.

8. Kampmann, M.-L., Schnell, I.B., Jensen, R.H., Axtner, J., Sander, A.F., Hansen, A.J., Bertelsen, M.F., Greenwood, A.D., Gilbert, M.T.P., and Wilting, A. (2017). Leeches as a source of mammalian viral DNA and RNA—a study in medicinal leeches. Eur. J. Wildl. Res. *63*, 36.

9. Kaczensky, P., Lkhagvasuren, B., Pereladova, O., Hemami, M., and Bouskila, A. (2015). Equus hemionus. The IUCN Red List of Threatened Species 2015: e. T7951A45171204.

10. Löhr, C.V., Juan-Sallés, C., Rosas-Rosas, A., García, A.P., Garner, M.M., and Teifke, J.P. (2005). Sarcoids in captive zebras (Equus burchellii): association with bovine papillomavirus type 1 infection. J. Zoo Wildl. Med. *36*, 74–81.

11. van Dyk, E., Oosthuizen, M.C., Bosman, A.-M., Nel, P.J., Zimmerman, D., and Venter, E.H. (2009). Detection of bovine papillomavirus DNA in sarcoid-affected and healthy free-roaming zebra (Equus zebra) populations in South Africa. J. Virol. Methods *158*, 141–151.

12. Johnson, J., Howard, K., Wilson, A., Ward, M., Gilbert, G.L., and Degeling, C. (2019). Public preferences for One Health approaches to emerging infectious diseases: a discrete choice experiment. Soc. Sci. Med. *228*, 164–171.

13. Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., and Daszak, P. (2008). Global trends in emerging infectious diseases. Nature *451*, 990–993.

14. Pruvot, M., Khammavong, K., Milavong, P., Philavong, C., Reinharz, D., Mayxay, M., Rattanavong, S., Horwood, P., Dussart, P., Douangngeun, B., *et al.* (2019). Toward a quantification of risks at the nexus of conservation and health: The case of bushmeat markets in Lao PDR. Sci. Total Environ. *676*, 732–745.

15. Swift, L., Hunter, P.R., Lees, A.C., and Bell, D.J. (2007). Wildlife trade and the emergence of infectious diseases. EcoHealth *4*, 25.

16. Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., *et al.* (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature, 1–4.

17. Peeri, N.C., Shrestha, N., Rahman, M.S., Zaki, R., Tan, Z., Bibi, S., Baghbanzadeh, M., Aghamohammadi, N., Zhang, W., and Haque, U. (2020). The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? Int. J. Epidemiol.

18. Drosten, C., Günther, S., Preiser, W., Van Der Werf, S., Brodt, H.-R., Becker, S., Rabenau, H., Panning, M., Kolesnikova, L., Fouchier, R.A., *et al.* (2003). Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N. Engl. J. Med. *348*, 1967–1976.

19. Baize, S., Pannetier, D., Oestereich, L., Rieger, T., Koivogui, L., Magassouba, N., Soropogui, B., Sow, M.S., Keïta, S., De Clerck, H., *et al.* (2014). Emergence of Zaire Ebola virus disease in Guinea. N. Engl. J. Med. *371*, 1418–1425.

20. Sharp, P.M., and Hahn, B.H. (2011). Origins of HIV and the AIDS pandemic. Cold Spring Harb. Perspect. Med. *1*, a006841.

21. Hoye, B.J., Munster, V.J., Nishiura, H., Klaassen, M., and Fouchier, R.A. (2010). Surveillance of wild birds for avian influenza virus. Emerg. Infect. Dis. *16*, 1827.

22. Wilking, H., Ziller, M., Staubach, C., Globig, A., Harder, T.C., and Conraths, F.J. (2009). Chances and limitations of wild bird monitoring for the avian influenza virus H5N1—detection of pathogens highly mobile in time and space. PLoS One *4*.

23. Śmietanka, K., Woźniakowski, G., Kozak, E., Niemczuk, K., Frączyk, M., Bocian, \Lukasz, Kowalczyk, A., and Pejsak, Z. (2016). African swine fever epidemic, Poland, 2014–2015. Emerg. Infect. Dis. *22*, 1201.

24. Bitome-Essono, P.-Y., Ollomo, B., Arnathau, C., Durand, P., Mokoudoum, N.D., Yacka-Mouele, L., Okouga, A.-P., Boundenga, L., Mve-Ondo, B., Obame-Nkoghe, J., *et al.* (2017). Tracking zoonotic pathogens using blood-sucking flies as' flying syringes'. elife *6*, e22069.

25. Nasi, R., Taber, A., and Van Vliet, N. (2011). Empty forests, empty stomachs? Bushmeat and livelihoods in the Congo and Amazon Basins. Int. For. Rev. *13*, 355–368.

26. Schnell, I.B., Bohmann, K., Schultze, S.E., Richter, S.R., Murray, D.C., Sinding, M.-H.S., Bass, D., Cadle, J.E., Campbell, M.J., Dolch, R., *et al.* (2018). Debugging diversity–a pan-continental exploration of the potential of terrestrial blood-feeding leeches as a vertebrate monitoring tool. Mol. Ecol. Resour. *18*, 1282–1298.

27. Abrams, J.F., Hörig, L.A., Brozovic, R., Axtner, J., Crampton-Platt, A., Mohamed, A., Wong, S.T., Sollmann, R., Yu, D.W., and Wilting, A. (2019). Shifting up a gear with iDNA: From mammal detection events to standardised surveys. J. Appl. Ecol. *56*, 1637–1648.

28. Ng, T.F.F., Willner, D.L., Lim, Y.W., Schmieder, R., Chau, B., Nilsson, C., Anthony, S., Ruan, Y., Rohwer, F., and Breitbart, M. (2011). Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. PloS One *6*.

29. Hoffmann, C., Stockhausen, M., Merkel, K., Calvignac-Spencer, S., and Leendertz, F.H. (2016). Assessing the feasibility of fly based surveillance of wildlife infectious diseases. Sci. Rep. *6*, 1–9.

30. Farkas, K., McDonald, J.E., Malham, S.K., and Jones, D.L. (2018). Two-step concentration of complex water samples for the detection of viruses. Methods Protoc. *1*, 35.

31. Axtner, J., Crampton-Platt, A., Hörig, L.A., Mohamed, A., Xu, C.C., Yu, D.W., and Wilting, A. (2019). An efficient and robust laboratory workflow and tetrapod database for larger scale environmental DNA studies. GigaScience *8*, giz029.

32. Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. *2010*, pdb–prot5448.

33. Alfano, N., Courtiol, A., Vielgrader, H., Timms, P., Roca, A.L., and Greenwood, A.D. (2015). Variation in koala microbiomes within and between individuals: effect of body region and captivity status. Sci. Rep. *5*, 10189.

34. Wang, D., Coscoy, L., Zylberberg, M., Avila, P.C., Boushey, H.A., Ganem, D., and DeRisi, J.L. (2002). Microarray-based detection and genotyping of viral pathogens. Proc. Natl. Acad. Sci. *99*, 15687–15692.

35. Yozwiak, N.L., Skewes-Cox, P., Stenglein, M.D., Balmaseda, A., Harris, E., and DeRisi, J.L. (2012). Virus identification in unknown tropical febrile illness cases using deep sequencing. PLoS Negl. Trop. Dis. *6*.

36. Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210.

37. Langmead, B., and Salzberg, S. (2012). Fast gapped-read alignment with bowtie 2 Nat Methods 9 (4): 357–359. pmid: 22388286 View Article PubMed (NCBI).

38. Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics *28*, 3211–3217.

39. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., *et al.* (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477.

40. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.* (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. *29*, 644.

41. Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

42. Vilsker, M., Moosa, Y., Nooij, S., Fonseca, V., Ghysens, Y., Dumon, K., Pauwels, R., Alcantara, L.C., Vanden Eynden, E., Vandamme, A.-M., *et al.* (2019). Genome Detective: an automated system for virus identification from high-throughput sequencing data. Bioinformatics *35*, 871–873.

43. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., *et al.* (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647–1649.

44. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

45. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

46. Somervuo, P., Koskela, S., Pennanen, J., Henrik Nilsson, R., and Ovaskainen, O. (2016). Unbiased probabilistic taxonomic classification for DNA barcoding. Bioinformatics *32*, 2920–2927.

47. Dayaram, A., Franz, M., Schattschneider, A., Damiani, A.M., Bischofberger, S., Osterrieder, N., and Greenwood, A.D. (2017). Long term stability and infectivity of herpesviruses in water. Sci. Rep. *7*, 46559.

48. Palmarini, M., Datta, S., Omid, R., Murgia, C., and Fan, H. (2000). The long terminal repeat of Jaagsiekte sheep retrovirus is preferentially active in differentiated epithelial cells of the lungs. J. Virol. *74*, 5776–5787.

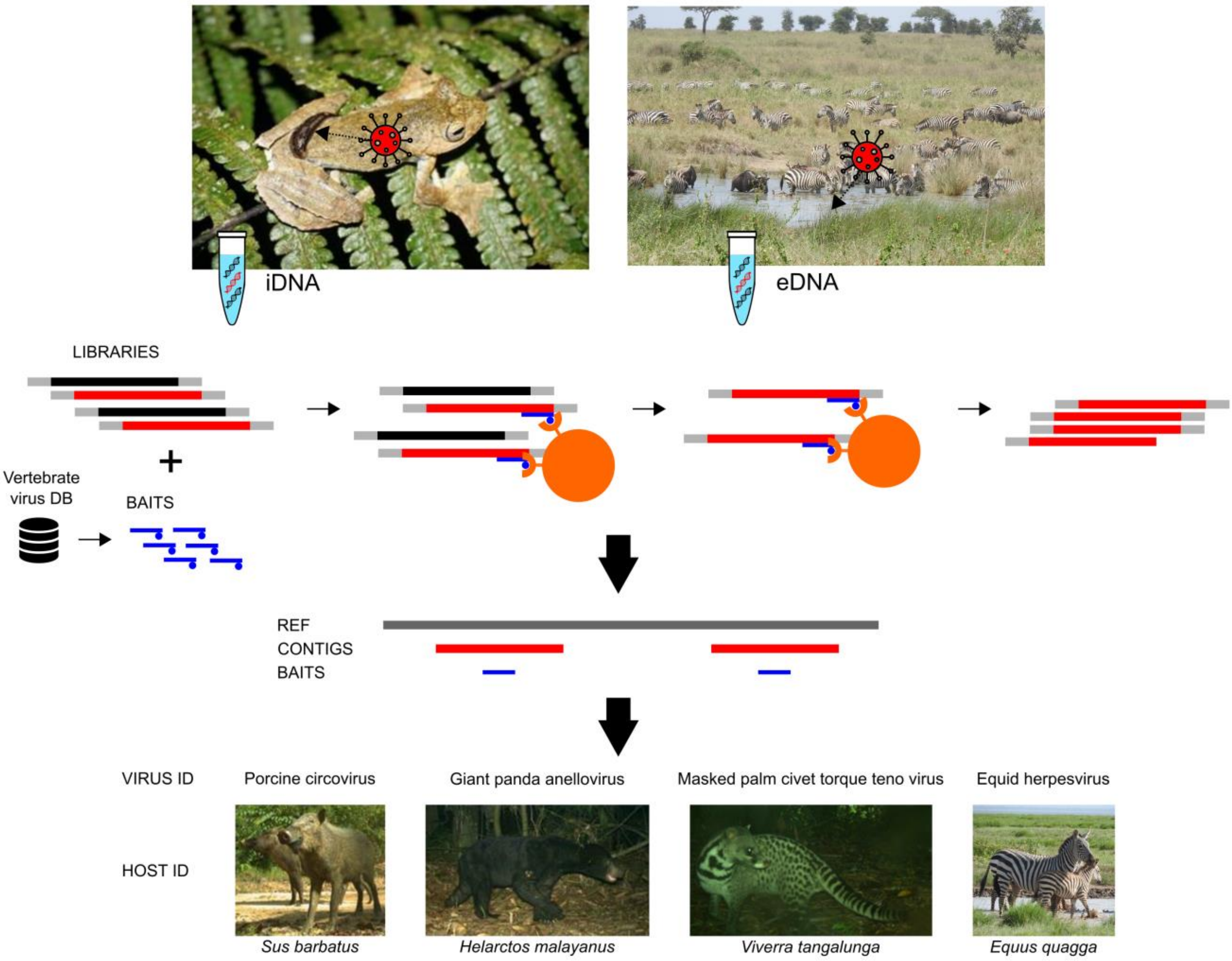

Figure 1: Viral screening of vertebrate viruses from leech iDNA and waterholes eDNA using RNA olignonucleotide based hybridization capture. In the upper panel the left photo shows a leech feeding on a frog in a rainforest of Vietnam (courtesy Andrew Tilker; Leibniz-IZW) and the right photo shows an African waterhole in Tanzania (courtesy Peter Seeber; Leibniz-IZW). The middle panel depicts the hybridization capture protocol. Briefly, Illumina libraries were produced from reverse transcribed RNA and DNA from leech bloodmeals or from waterhole surface water and sediments. Biotinylated viral RNA baits were hybridized to the libraries and non-target DNA was washed away. The remaining DNA was sequenced, reads assembled into contigs and mapped to reference viral genomes. These contigs were further analyzed to determine viral identity. Viral identity was paired with host identity determined either by mammalian metabarcoding of the leech samples, or by observation of waterhole usage.

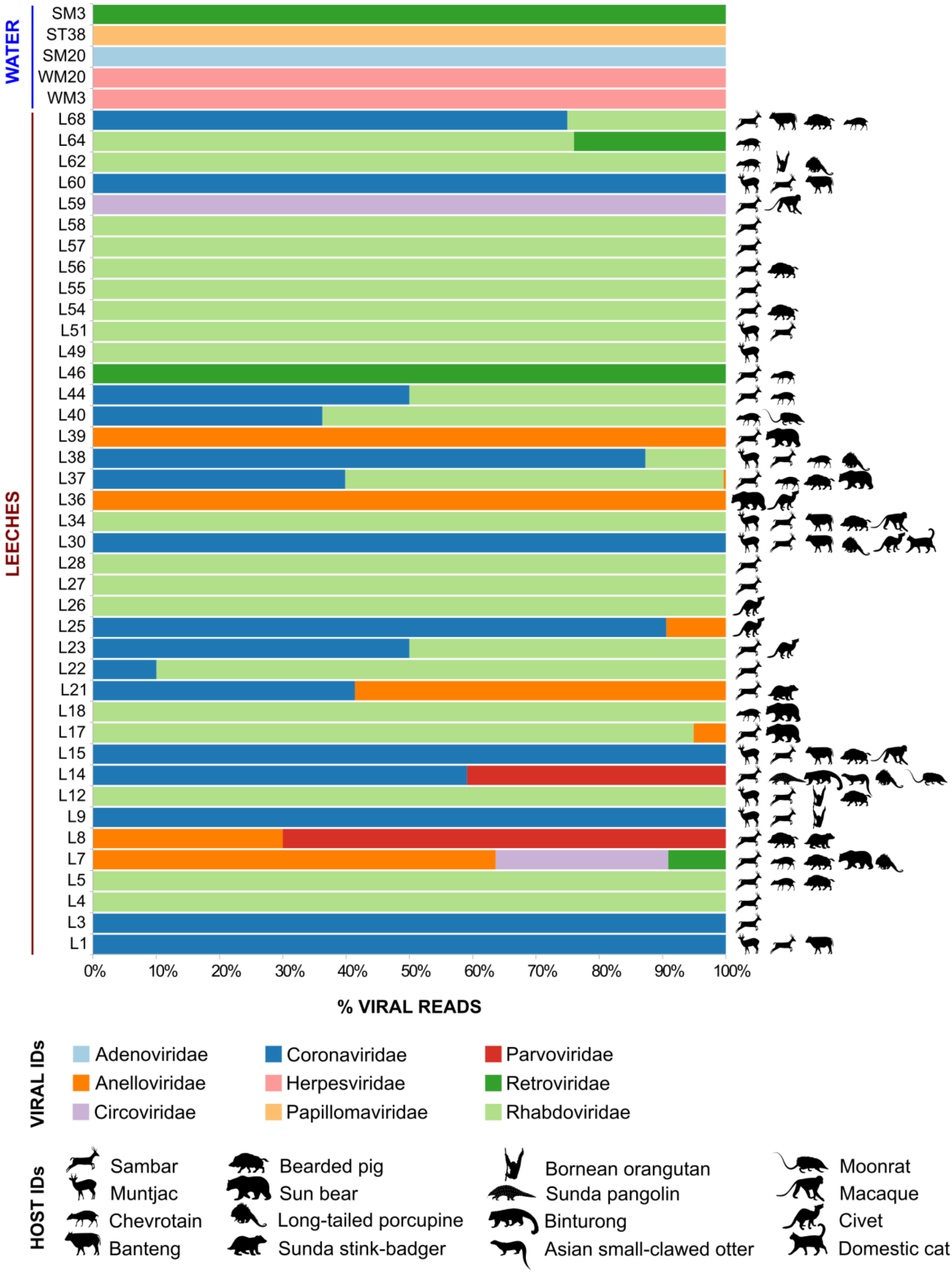

Figure 2: Relative abundance of viruses from each family, shown as the percentage of the total number of viral reads in each leech and waterhole sample. In the sample names, S stands for sediment, W for water, T for Tanzania, M for Mongolia and L for leeches. The leech host assignment for each leech sample is shown on the right (see Suppl. Tab. 1 for further details).