High-Depth Viral Blood Metagenomics Reveals Extensive Anellovirus Diversity in Healthy Humans

María Cebriá-Mendoza
Universitat de València
Cristina Arbona
Centro de transfusión de la Comunidad Valenciana
Luís Larrea
Centro de transfusión de la Comunidad Valenciana
Wladimiro Díaz
Universitat de València
Vicente Arnau
Universitat de València
Carlos Peña
Spanish National Research Council
Juan Bou
Universitat de València
Rafael Sanjuán
Universitat de València
José Cuevas (cuevast@uv.es)
Universitat de València

Research Article

Keywords: blood virome, anellovirus, virus discovery, metagenomics

Posted Date: December 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-115092/v1

Version of Record: A version of this preprint was published at Scientific Reports on March 25th, 2021. See the published version at https://doi.org/10.1038/s41598-021-86427-4.

High-depth viral blood metagenomics reveals extensive anellovirus diversity in healthy humans

María Cebriá-Mendoza[a], Cristina Arbona[b], Luís Larrea[b], Wladimiro Díaz[a,c], Vicente Arnau[a,c], Carlos Peña[a], Juan Vicente Bou[a], Rafael Sanjuán[a,d], José M. Cuevas[a,d*]

[a]Institute for Integrative Systems Biology (I2SysBio), Universitat de València-CSIC, València, Spain.

[b]Centro de Transfusión de la Comunidad Valenciana, Valencia, Spain.

[c]Department of Informatics, Universitat de València, València, Spain

[d]Department of Genetics, Universitat de València, València, Spain

*Corresponding author: José Manuel Cuevas, Institute for Integrative Systems Biology (I2SysBio), Universitat de València-CSIC, 46980 Paterna, València, Spain. Phone: 34 963 543 667, FAX 34 963 543 670, e-mail: cuevast@uv.es

Running title: Anellovirus diversity in humans

4079 words in the main text, 156 words in the summary, 1 table, 4 figures, 10 supplementary tables and 2 supplementary figures.

Footnote page:

Competing interests: The authors declare no competing interests.

Summary

Human blood metagenomics has revealed the presence of different types of viruses in apparently healthy subjects. By far, anelloviruses constitute the viral family that is more frequently found in human blood, although amplification biases and contaminations pose a major challenge in this field. To investigate this further, we subjected pooled plasma samples from 120 healthy donors in Spain to high-speed centrifugation, RNA and DNA extraction, random amplification, and massive parallel sequencing. Our results confirm the abundance of anelloviruses in such samples, which represented >98% of the total viral sequence reads obtained. We assembled 114 different viral genomes belonging to this family, revealing remarkable diversity. Phylogenetic analysis of ORF1 suggested 28 potentially novel anellovirus species, 24 of which were validated by Sanger sequencing to discard artifacts. These findings underscore the importance of implementing more efficient purification procedures that enrich the viral fraction as an essential step in virome studies and question the suggested pathological role of anelloviruses.

Keywords: blood virome; anellovirus; virus discovery; metagenomics

Introduction

The increasing amount of information provided by metagenomics has accelerated the discovery of novel viruses, showing overwhelming viral diversity at all levels[1]. Viral metagenomics has been used to identify viral agents causing disease outbreaks or associated with specific symptoms[2,3], to study the virosphere diversity[4–6], and to address specific aspects of viral evolution[7,8]. Many of the newly discovered viruses are not associated with any disease and are consequently called "orphans"[9]. The family *Anelloviridae* provides the clearest example, since only one member of the genus *Gyrovirus* has been confirmed to cause disease in chickens[10], despite an increasing number of anelloviruses being discovered in wild and domestic animals[11–15]. Three genera are known to produce chronic infections in humans: torque teno virus (TTV, *Alphatorquevirus*)[16], torque teno mini virus (TTMV, *Betatorquevirus*), and torque teno midi virus (TTMDV, *Gammatorquevirus*). Indeed, anelloviruses constitute the most prevalent human-infective viruses[17].

Little is known about the biology of anelloviruses because of the lack of appropriate cell cultures and animal models. However, it has been established that human anelloviruses are distributed worldwide and are frequently present in blood, feces, semen and urine[18]. Since the discovery of the first anellovirus[19], the diversity of this family is constantly increasing as new members are identified. The family *Anelloviridae* currently encompasses fourteen genera, and the ICTV has subdivided TTV, TTMV, and TTMDV into 29, 12 and 15 species, respectively. Taxonomic classification is currently based on the analysis of the entire ORF1 nucleotide sequence, with pairwise nucleotide sequence identity cut-off values of 35% and 56% to define a species and a genus, respectively[20].

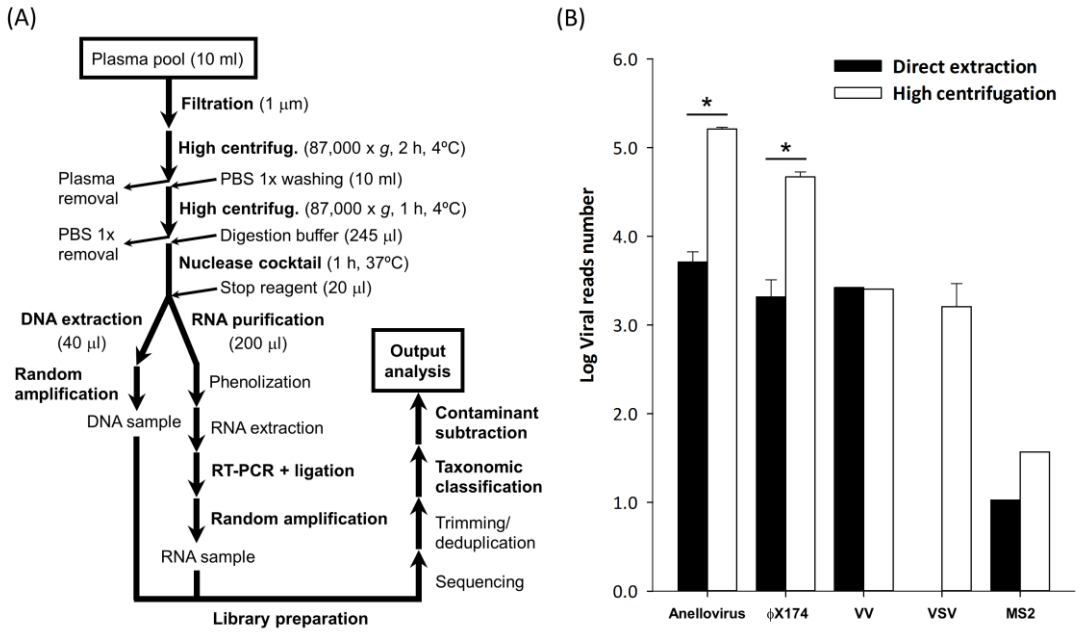
Studies analyzing the blood virome of apparently healthy individuals have also revealed the presence of unknown viruses[21–23], which is particularly relevant when considering blood transfusions or organ transplantation[24]. Anelloviruses occupy the largest fraction of the blood virome[25]. Based on previous studies[26–28], we have used a protocol involving high-speed centrifugation, random RNA and DNA amplification, and massive sequencing of 120 pooled-plasma samples from blood donors in order to characterize viral diversity. The multiple displacement amplification (MDA)[29] method was used for random amplification, which preferentially amplifies circular single stranded DNA but has been successfully used to detect RNA viruses in biological samples[29]. Additionally, since contaminant nucleic acids potentially causing misleading results were expected to be present along the purification protocol[30], three blank controls were also used for eventual subtraction of the identified taxons.

Results

Strategy and overall sequence output

The protocol used in this study intended to enrich the viral fraction and different experimental combinations involving filtration and centrifugation steps were initially tested as explained below (Fig. 1A). Subsequent nuclease digestion for removing free nucleic acids was performed before independent viral DNA and RNA extraction, followed by random amplification and library preparation. Sequencing results were then taxonomically classified and taxons present in blank controls were subtracted. Since filtration and washing steps should select for viral particles, the vast majority of human and bacterial nucleic acids are expected to be removed or eventually subtracted in subsequent bioinformatics analysis.

Figure 1. Experimental and bioinformatics workflow (A) and comparison between viral reads recovered with direct extraction from plasma and the protocol involving initial high centrifugation (B). Main steps at panel A are marked in bold (See details in Methods section). For panel B, error bars indicate standard error of the mean (SEM, *n* = 2 replicates). Asterisk indicates the statistical significance of a log-scale *t*-test analyzing the efficiency of the purification protocols ($*P < 0.001$). Viral reads per million of total reads are used. For VV, the only indicated value for each treatment was obtained with 1 μm pore size filtration.



We initially set out to compare viral recovery efficiency obtained by directly extracting nucleic acids from plasma or by performing a high-speed centrifugation step first. For this, we spiked a plasma pool, including ten individual plasma samples, with bacteriophages ϕX174 (non-enveloped, circular single-stranded DNA virus) and MS2 (non-enveloped, linear single-stranded RNA virus), vaccinia virus (VV, large enveloped, linear double-stranded DNA virus), and vesicular stomatitis virus (VSV, enveloped, linear single-stranded RNA virus), and used them for massive parallel Illumina sequencing (see details in Methods section). In this pilot study, we also analyzed anelloviruses (non-enveloped, circular single-stranded DNA viruses), since these are frequently found in blood. Two technical replicates differing in the pore size used at the initial filtration step (0.45 vs 1.0 μm) were performed, although this difference is only expected to affect large viruses such as VV[31]. Indeed, samples initially filtered with the largest pore size yielded thousands of VV reads, while VV was not detected in one of the samples filtered with the smaller pore size (Fig. 1B and Supplementary Table S1). When checking the presence of circular DNA viruses (anelloviruses and ϕX174), clear increases of viral recovery efficiency ranging one or two orders of magnitude were observed in the protocol involving a high-speed centrifugation step (Fig. 1B; *t*-test: $P < 0.001$ for both viruses). VSV reads were not detected in the direct extraction treatment, whereas thousands of reads were recovered when using high-speed centrifugation. For MS2, viral reads were detected in a single replicate from each treatment and no clear conclusions could be drawn, although this could be accounted for by the low amount initially added for this virus. Since our results indicated that a high-speed centrifugation step substantially increased the recovery of circular DNA viruses and VSV, this approach was used thereafter in combination with an initial filtration step using a 1 μm pore size to avoid potential loss of large viruses. The sample obtained from ten pooled plasmas in this pilot study using the

1 μm filter and high-speed centrifugation was named pool 1 (P1). The samples subsequently obtained using these conditions, each including ten individual plasma samples, were named accordingly (P2-P12).

Since the purification protocol might carry over residual amounts of nucleic acids, it was essential to introduce blank controls to evaluate contamination risk. The reads obtained in these controls were used for taxonomic classification and subtraction of these potential contaminants from real samples using Centrifuge[32] and Recentrifuge[33] software, respectively. The discarded reads were mainly of human, bacterial, and viral origin (Table 1 and Supplementary Table S2). We noticed that ambiguities in the taxonomical classification of reads were not properly handled by Recentrifuge, limiting our ability to remove potential contaminations corresponding to phylogenetically unclassified reads. Reassuringly, the total fraction of viral reads increased from 57.7% to 90.8% after the subtraction step (Fig. 2A). The total fraction of bacterial reads after subtraction dropped from 37.9% to 9.2%. As expected, human reads were removed by Recentrifuge. The non-removed bacterial reads encompassed 32 phyla (Fig. 2B and Supplementary Table S3), including Firmicutes (35.7%), Proteobacteria (31.9%), Bacteroidetes (17.4%), Cyanobacteria (4.9%) and Actinobacteria (3.3%). The relative abundances of these phyla is consistent with previous blood microbiome studies[34], suggesting that these sequences may correspond to residual amounts of DNA that survived our virus-enrichment protocol. Alternatively, these could be contaminants that were not removed computationally.

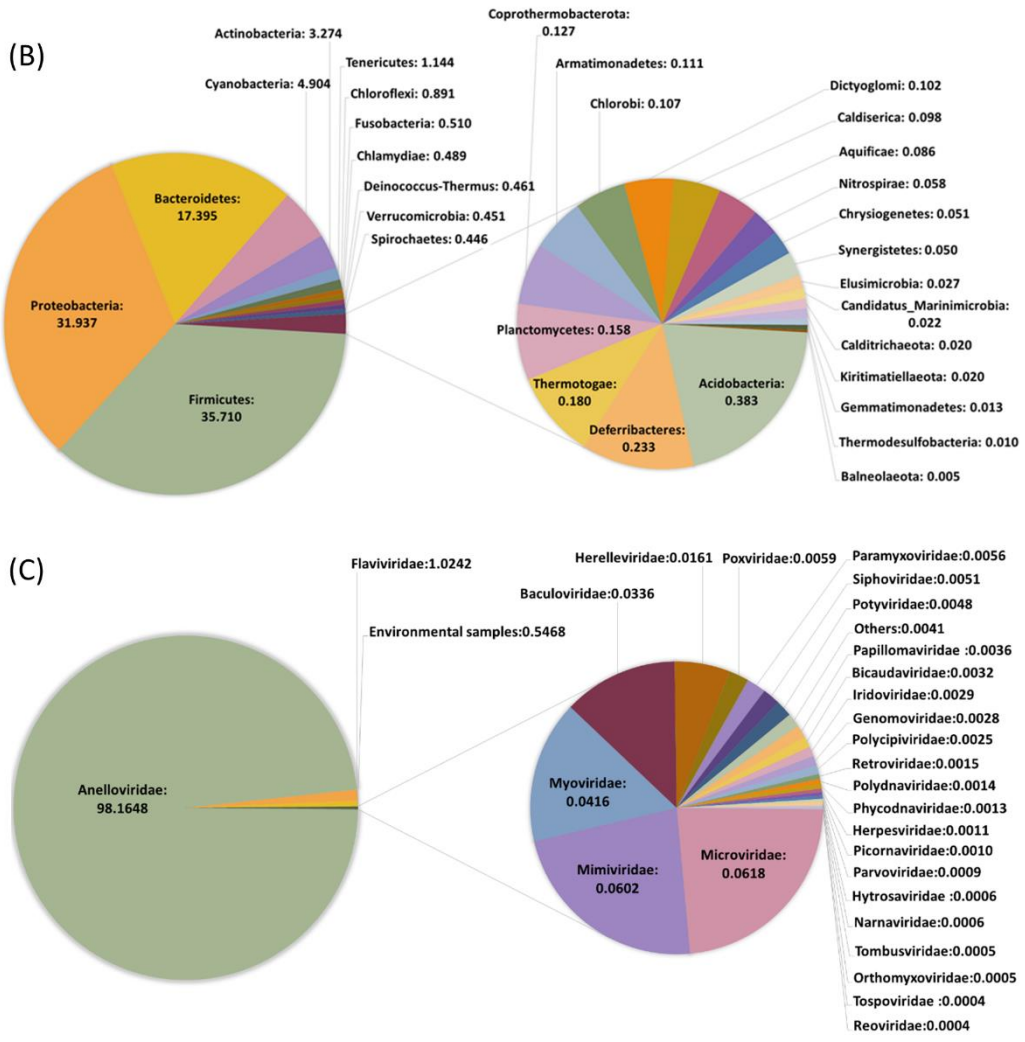
Table 1. Summary of Recentrifuge results for the 12 pools analysed. For each pool, the total number of reads passing Recentrifuge analyses, and those classified as bacterial, anellovirus and other viruses, are indicated. Last column shows the number of anellovirus contigs over 1.5 kb obtained in the assembling step. For the sake of clarity, viral reads from spiked viruses are excluded from counts.

Pool	# total reads	Bacterial reads	Anellovirus reads	Other viruses	Anellovirus contigs
P1	183,836	3,517	139,733	627	25
P2	123,617	16,815	82,249	2,125	9
P3	123,390	5,446	106,777	530	9
P4	131,804	7,838	107,249	719	4
P5	77,249	35,525	9,512	1,179	2
P6	47,957	23,741	5,641	46	3
P7	336,968	8,157	304,349	2,945	22
P8	195,036	8,758	168,675	2,213	20
P9	509,882	8,927	461,043	16,485	6
P10	155,952	16,412	117,288	1,739	8
P11	76,143	16,053	22,376	228	2
P12	73,048	9,804	40,339	403	4

Figure 2. Summary of bioinformatics subtraction for human, bacterial and viral groups (A) and description of the microbiome (B) and the virome (C) characterized in this study. Classification is shown for bacteria and viruses at phylum and family level, respectively. Frequencies were obtained excluding spiked virus contribution in A and C panels.

(A)

	%reads		
	Human	Bacterial	Viral
Centrifuge	4.4	37.9	57.7
Recentrifuge	0.0	9.2	90.8



Our samples contained sequences from 28 different viral families (Fig. 2C and Supplementary Table S4), but with a clear dominance of the *Anelloviridae* family, which represented more than 98% of the total fraction. The second most abundant family was *Flaviviridae* (1.0%), although most reads corresponded to a GB virus C detected in pool 9 (16258 reads; genome coverage >95%, depth 191x). This finding confirmed that our protocol was also efficient for RNA virus recovery. A third group encompassed 0.55% of the total fraction and was classified as "environmental samples" (NCBI: txid186616). The remaining viral families were detected at very

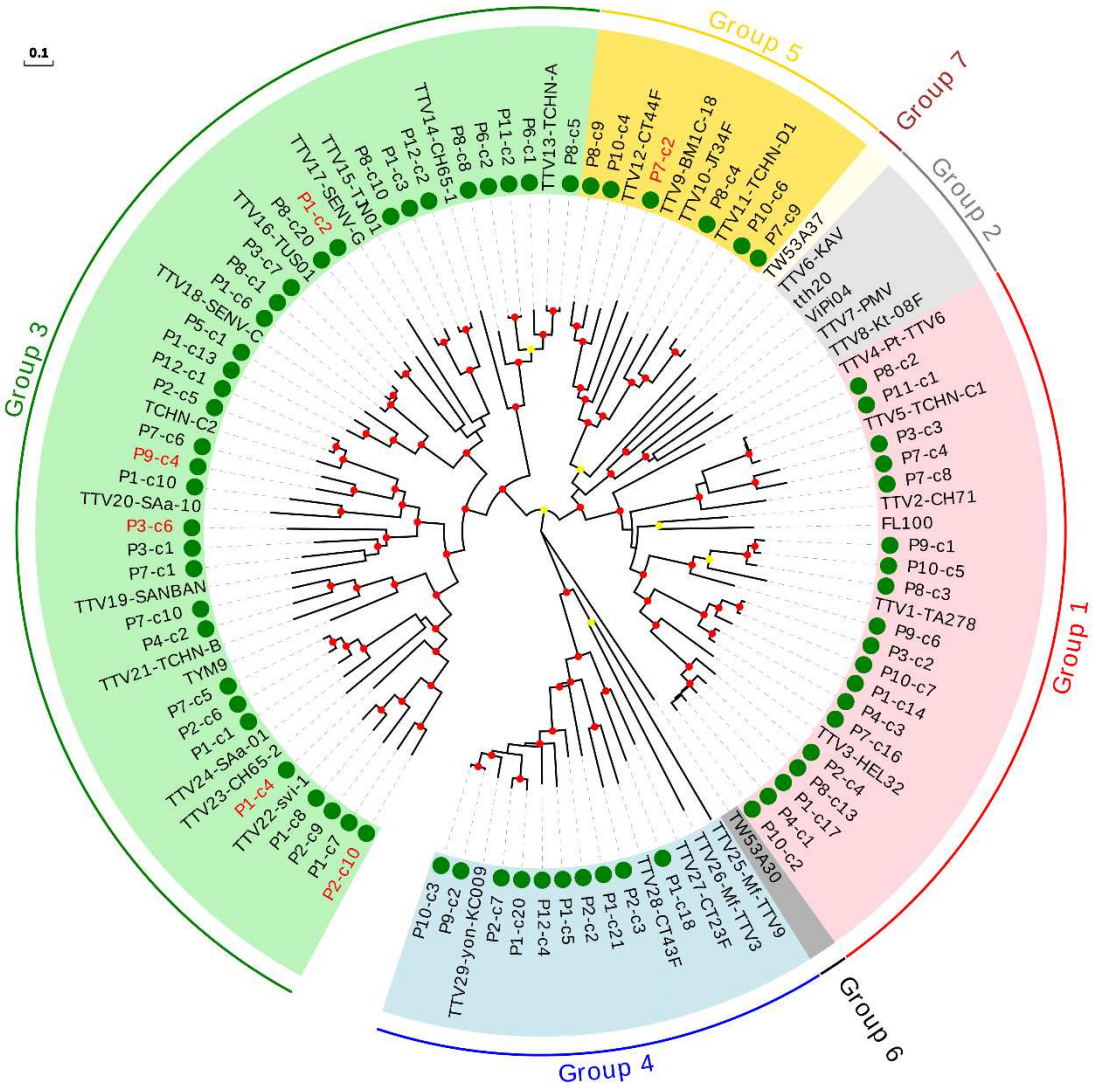
low frequencies, with read number ranging between 6 and 963. It is worth mentioning that reads assigned to the family *Circoviridae* were systematically subtracted by Recentrifuge. The detection of members from this family has been associated with contaminated reagents[30], which stresses the necessity of including appropriate controls.

Analysis of anelloviruses

For each of the 12 pools, we generated contigs from reads, and those larger than 1.5 kb were subsequently analyzed. Blast analyses showed that only a few contigs belonged to the viruses spiked in pools 1 and 2 or the abovementioned GB virus C, whereas 114 contigs corresponded to anelloviruses, of which 23 showed overlapping ends and could thus be considered as complete genomes (Table 1 and Supplementary Table S5). Additionally, there was a significantly positive correlation between the number of contigs and the total amount of anelloviral reads in each pool (Spearman's correlation: $\rho = 0.728$; $P = 0.004$). We used the ORF1 nucleotide sequence for phylogenetic analysis. Full-length ORF1 was obtained for all but eight out of the 114 contigs (93%). For a preliminary taxonomic classification, we constructed a phylogenetic tree including ORF1 from Genbank hominid sequences (Supplementary Table S6), which allowed assignment of our contigs as belonging to TTV, TTMV or TTMDV genera (68, 29, and 17 sequences, respectively; Supplementary Table S5 and Supplementary Fig. S1). From the 23 contigs considered as complete genomes, 22 and one belonged to TTMV and TTMDV genera, respectively. Assembly efficiency was strongly affected by GC-rich regions present in anelloviruses, but these regions are shorter in TTMV genus[35], which can facilitate full-length genomes completion. This also explained why several contigs fell into the expected full-length genome size range but did not present terminal redundancy.

In order to aid visualization, phylogenetic trees were independently constructed for each genus, and only one representative genotype of each species was used, including some that are not currently accepted by ICTV. For the TTV genus, which has been postulated to consist of seven phylogenetic groups[16], the tree included our 68 new sequences as well as 36 previously described genotypes, each representing one known species (Fig. 3). This tree, along with divergence values, indicated that eight of our sequences could be considered as belonging to six novel species, whereas the remaining sequences clustered within 18 of the 36 previously known species (Supplementary Table S7). The number of our sequences assigned to each species was variable. For instance, four species clustered with only one of our sequences, whereas the species represented by genotypes TTV29-yon-KC009 and TTV3-HEL32 clustered with eight and ten of our sequences, respectively (Supplementary Fig. S2). This is in contrast with a previous study showing that TTV8 was the most quantitatively prevalent species in human blood[25], as TTV8 did not cluster with any of our sequences. We also found no sequences that clustered with species belonging to groups 2, 6, and 7. However, there was a significant positive correlation between the number of species included in each group and the number of newly described sequences, even when discarding data from the recently proposed groups 6 and 7, which consisted of a single species (Spearman's correlation coefficient; $\rho = 0.821$, $P = 0.044$).

Figure 3. Phylogenetic tree for the ORF1 including the representative genotypes from TTV genus. Sequences described in this study are marked with a green circle. Those sequences that could be considered as new species are labelled in red. 0.7-0.85 and 0.85-1.0 bootstrap value ranges are indicated with yellow and red circles, respectively. Scale bar indicates evolutionary distance in nucleotide substitutions per site.

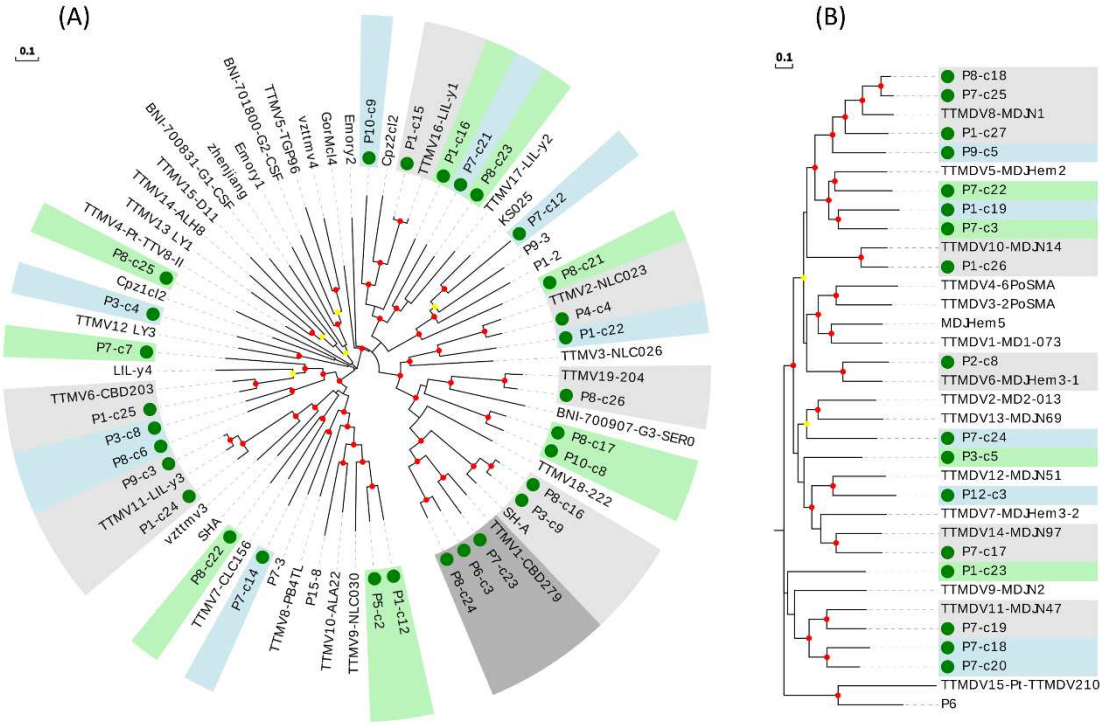


We then constructed a phylogenetic tree including 29 sequences from our study belonging to the TTMV genus and the 38 previously-described representative genotypes meeting the species demarcation criteria (Fig. 4A). Surprisingly, despite the smaller number of new TTMV sequences identified compared with the TTV genus (29 vs 68), most could be considered as novel species (Supplementary Table S8). In total, fifteen novel species were defined, three of which included two sequences, while the remaining 11 isolates clustered with seven of the 38 previously described species.

For the TTMDV genus, we identified 17 new sequences and used them to build a tree that also included 17 representative genotypes from known species. Similar to TTMV, we found that most new sequences are likely to correspond to novel species (Fig. 4B and Supplementary Table S9). Ten of our sequences defined 9 novel species, whereas the remaining 7 TTMDV sequences

clustered with five of the 17 representative genotypes belonging to previously described species.

Figure 4. Phylogenetic trees for the ORF1 including the representative genotypes from TTMV (A) and TTMDV (B) genera. Sequences described in this study are marked with a green circle. New species (including one or more new sequences) are indicated with background green or blue color in order to distinguish contiguous clusters. Clusters of representative species including new sequences are indicated with background light or dark grey colors in order to distinguish contiguous clusters. 0.7-0.85 and 0.85-1.0 bootstrap value ranges are indicated with yellow and red circles, respectively. Scale bar indicates evolutionary distance in nucleotide substitutions per site.



Assembly of massive sequencing reads could produce artifacts, eventually affecting the reliability of phylogenetic analysis. Since this possibility was particularly relevant for the assignation of new species, the DNA extracts from which we obtained 24 of the proposed new TTV, TTMV, and TTMDV species (4, 11, and 9 samples, respectively) were selected for reanalysis. These samples differed in sequencing coverage, ranging from 2x to 482x after assembly. For each, we performed sequence-specific PCR amplification and Sanger sequencing of the complete ORF1 (Supplementary Table S10). In all cases, Sanger sequencing confirmed the ORF1 sequences previously inferred by random amplification and Illumina sequencing, highlighting the reliability of our pipeline.

Discussion

Implementation of large-scale blood virome studies is a powerful tool for the early detection of human emergent viruses causing chronic infections or exhibiting long asymptomatic phases, although surveillance programs based on this approach have not been established widely. Our results show that using adequate controls is essential in these studies, since contaminations can lead to false positives[36,37]. In our study, we have used three negative controls throughout the experimental protocol, and taxons eventually identified in these controls have been

computationally subtracted from the samples. Since samples were initially filtered and used for digestion of free nucleic acids, we expected the non-viral fraction to be drastically reduced. Nevertheless, our data contained a significant fraction of bacterial and human reads, but these could be significantly reduced by bioinformatics subtraction. In our study, we have used MDA assay for random amplification, which can preferentially amplify circular single stranded DNA viruses[29], such as anelloviruses. This amplification bias can partially explain the overwhelming abundance of this family in our results, but sensitive detection of an RNA virus confirms the robustness of the proposed procedure.

Viral metagenomics should also benefit strongly from the implementation of procedures involving pre-amplification, purification, and enrichment steps as the one described here, since this increases sensitivity[27,28]. Indeed, this is supported by a recent study that analyzed anellovirus distribution in small mammals, in which sample purification involved sucrose gradient ultracentrifugation[11]. This study detected 11 potential novel species, and proposed the inclusion of two novel genera within the *Anelloviridae* family.

Together with previous studies, our result show that the diversity of anelloviruses is particularly remarkable in comparison with other viral families[9]. The fact that human and non-human primate isolates cluster phlylogenetically[38] suggests that anelloviruses are an ancient family, and that the genetic diversity of this family is the consequence of millions of years of evolution. It has recently been proposed that the increasing amount of viral sequences identified by metagenomics should be incorporated into the ICTV classification scheme[39]. This inclusion, which should require appropriate quality control, is important for obtaining a more realistic picture of viral global diversity. Although this proposal is particularly relevant for environmental samples, the ICTV picture for *Anelloviridae* does not reflect its continuously increasing diversity.

Most of the sequences detected in our study belong to the TTV genus, which has been more extensively studied than other anellovirus genera. Yet, potentially novel species were mainly found among TTMV and TTMDV genera. It is likely that the later are more difficult to detect in protocols lacking viral enrichment and, hence, remain more poorly characterized. As such, our results underscore the importance of using viral enrichment methods for the study of anellovirus diversity.

It has been proposed that anellovirus abundance in blood increases in immunosuppressed patients, as has been described in transplanted[40] and HIV-1 patients[41,42]. It has also been shown that anellovirus prevalence is lower in healthy subjects than in patients with common pathologies[43]. This has led to the suggestion that viral load could be used as a health biomarker in patients with chronic conditions, or even in people without known pathologies[44,45]. TTVs have also been postulated as biomarkers for anthropogenic pollution[46], graft rejection[25], and immune status[40]. However, cause-effect relationships between TTV load and health status need to be better clarified.

The prevalence of TTMV and TTMDV is markedly lower than that of TTV[25,43]. Overall, apart from some indirect evidence, viruses from the TTDMV genus have not been associated with pathologies[51]. In contrast, many of the recently described anellovirus species belonging to the TTMV genus have been associated to specific pathologies[47–50]. As a note of caution, associations between the presence of a virus and a pathological condition does not necessarily prove causality. As indicated above, anellovirus load could be a consequence of immune status. A lower load in healthier individuals could limit viral detection, leading to a statistical (but causal) association between the presence of a given virus and certain diseases. An illustrative example

of this possibility is given by genogroup 2 from TTV, which has been detected at a very low frequency in the healthy population[52]. Sequencing and qPCR studies, including our results, have shown that genogroup 2 is absent or detected at low frequencies in healthy donors[53–55], sporadically absent in transplanted patients[56], and detected at higher frequencies in immunosuppressed patients [25,53,55,57]. In addition, it has also been shown that TTV viral load increases with the number of TTV genogroups simultaneously infecting a patient[52,53], and that transplantation influences genogroup distribution[53].

The metagenomics era has led to a new ecological perspective in virology, which avoids considering viruses necessarily as disease-causing pathogens[58]. Instead, viruses are regarded as integral components of ecosystems that can sporadically cause emerging diseases but also can be beneficial to their hosts[59,60]. Human anelloviruses, and probably most members of this family, seem to be essentially innocuous[17]. Indeed, potentially beneficial effects on human health have been suggested[9]. For instance, infection of newborns[61] could promote the development and maturation of the immune system[17]. Besides, the detection of the same type of TTV in samples collected 16 years apart support the theory that people can remain chronically infected[62]. These results are in agreement with a long history of coevolution between the virus and the host, eventually leading to commensal or even mutualistic relationships.

Methods

Sample collection

A total of 120 plasma samples from healthy donors were collected from the Centro de Transfusión de la Comunidad Valenciana (Valencia, Spain) from September 15, 2018 to March 30, 2019. All samples were stored at -80^{o}C until use. All subjects gave written consent in accordance with the declaration of Helsinki. The protocol was approved by the University of Valencia ethics committee (IRB No. H1489496487993). Plasma samples were divided into 12 heterogeneous pools in age and gender (each pool included ten samples).

DNA/RNA extraction and amplification

For a pilot study, an initial pool of 10 samples was obtained (P1 pool) by mixing 2.5 mL of plasma from each sample into one tube (25 mL total). To assess viral recovery, we spiked this pool with 10^3 PFU/mL of ϕX174, vaccinia virus (VV) and MS2, and 10^4 PFU/mL of vesicular stomatitis virus (VSV). Half of the total volume was then filtered through a 0.45 μm or a 1.0 μm filter to remove cells and other non-viral particles. Since this different filtration is only expected to compromise the detection of big viruses[31], both filtered fractions could be considered technical replicates for all spiked viruses except VV. From each fraction, 1 mL was used to extract nucleic acids with the QIAAMP Ultra Sens Virus Kit (Qiagen) following the manufacturer's instructions. DNA was amplified from the final elution with the TruePrime WGA kit (Sygnis), whereas RNA from half of the final elute volume was cleaned with TRIzol LS reagent (Invitrogen), extracted with the QIAamp Viral RNA Mini kit and amplified using the QuantiTect Whole Transcriptome kit (Qiagen), which includes a ligation step following reverse transcription. In parallel, 10 mL from each filtered fraction was subject to high-speed centrifugation (87,000G, 2 h, 4^{o}C), washed with PBS 1X (87,000G, 1 h, 4^{o}C) and resuspended in 245 μL 1X digestion buffer (Turbo DNA Free kit, Ambion). Then 5 μL of Turbo DNase, 2 μL of Benzonase (Sigma) and 2 μL of micrococcal nuclease (NEB) were added to the sample to remove unprotected nucleic acids. After incubation (1h, 37^{o}C), 20 μL of stop reagent was added, following the manufacturer's instructions. Then, 240 μL supernatant was transferred to a new tube and split into two fractions: 200 μL fraction was used

for RNA extraction and final amplification as previously described, and 40 µL fraction was used for DNA extraction with the QIAamp Viral RNA Mini kit and amplification with the TruePrime WGA kit.

For the other eleven groups (P2-P12 pools), mixes were done adding 1 mL plasma from each sample. As a control, 5 x 10^2, 10^3 and 10^4 PFU of ϕX174, MS2 and VSV were added, respectively, to pool 2. Three blank samples starting the whole extraction protocol from 10 mL PBS 1X were used for subtraction of potentially contaminant taxons.

Massive parallel sequencing

For the pilot study, the 8 amplification products obtained (2 replicates x 2 extraction methods, direct extraction versus high-speed centrifugation x 2 types of products, DNA/RNA) were used for library preparation using Nextera XT DNA library preparation kit (Illumina) and sequenced using a MiSeq device. For the rest of the pools, DNA and RNA amplification products were mixed in equimolar concentration before library preparation and sequenced in a NextSeq device. The raw sequence reads from the metagenomic libraries were deposited in the Short Read Archive of GenBank database under accession number XXX.

Sequence analysis

Sequence data was quality checked using FastQC v0.11.8 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC v1.8[63]. Reads were first deduplicated using clampify.sh and then quality filtered using bbduk.sh, both from BBTools suite v38.68[64]. A quality trimming threshold of 20 was used and reads below 70 nucleotides in length were removed from the dataset.

The metagenomics analysis was carried out using the Centrifuge software package[32] version 1.0.4. A customized database was generated from the NCBI nt database downloaded in June 2019. The Centrifuge download tool was used to incorporate archaea, viruses, bacteria and fungi genomes from the NCBI RefSeq database. Finally, draftGenomes[65] was used to supplement the database with the SMS sequences in the NCBI WGS database belonging to viral taxa. Centrifuge results were postprocessed for contaminant removal and analyzed with Recentrifuge[33] version 1.1.0 using a minscore of 22. Assembly was performed with SPAdes[66].

Putative open reading frames were identified using ORF Finder (https://www.ncbi.nlm.nih.gov/orffinder/).

Phylogenetic analysis

To study phylogenetic relationships in *Anelloviridae* family, ORF1 sequences from hominid TTV, TTMV, and TTMDV isolates available from Genbank by February 2020 were downloaded (Supplementary Table S6). Sequence alignment (on the basis of the amino acid sequences) was performed with MUSCLE[67] and subsequent phylogenetic inference using nucleotide sequences was conducted with the maximum likelihood (ML) method implemented in MEGA version X[68]. Analysis were performed under the best fit nucleotide substitution model identified as $\mathrm{GTR} + \Gamma + \mathrm{I}$ using Akaike information criterion as the model selection framework in MEGA version X[68]. The reliability of the phylogenetic results was assessed using 1000 bootstrap replicates. The final trees were annotated with EvolView[69]. Species demarcation was performed by checking nucleotide pairwise identity matrices obtained independently for each genus.

Sanger sequencing

Sequence data obtained from assembled contigs for several anelloviruses were used to design primers amplifying the complete ORF1. Then, 25 μL PCR reactions were performed adding 1 μL DNA, Phusion High-fidelity DNA polymerase (ThermoFisher Scientific) and GC buffer using specific annealing conditions for each amplification product. PCR and additional internal primers were used for Sanger sequencing (Supplementary Table S10).

Bibliography

1. Koonin, E. V. & Dolja, V. V. Metaviromics: a tectonic shift in understanding virus evolution. *Virus Res.* **246**, A1–A3 (2018).

2. Delwart, E. Animal virus discovery : improving animal health , understanding zoonoses , and opportunities for vaccine development. *Curr. Opin. Virol.* **2**, 344–352 (2012).

3. Kapoor, A. *et al.* A highly prevalent and genetically diversified Picornaviridae genus in South Asian children. *Proc.Natl.Acad.Sci.U.S.A* **105**, 20482–20487 (2008).

4. Chow, C.-E. T. & Suttle, C. A. Biogeography of Viruses in the Sea. *Annu. Rev. Virol.* **2**, 41–66 (2015).

5. Paez-Espino, D. *et al.* Uncovering Earth's Virome. *Nature* **536**, 425–430 (2016).

6. Zhang, Y.-Z., Shi, M. & Holmes, E. C. Using Metagenomics to Characterize an Expanding Virosphere. *Cell* **172**, 1168–1172 (2018).

7. Li, C. *et al.* Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *Elife* **4**, e05378 (2015).

8. Shi, M. *et al.* The evolutionary history of vertebrate RNA viruses. *Nature* **556**, 197–202 (2018).

9. Kaczorowska, J. & Hoek, L. Van Der. Human anelloviruses : diverse , omnipresent and commensal members of the virome. *FEMS Microbiol. Rev.* 305–313 (2020). doi:10.1093/femsre/fuaa007

10. Li, Y. *et al.* Genomic Characterization of Recent Chicken Anemia Virus Isolates in. *Front. Microbiol.* **8**, 401 (2017).

11. De Souza, W. M. *et al.* Discovery of novel anelloviruses in small mammals expands the host range and diversity of the Anelloviridae. *Virology* **514**, 9–17 (2018).

12. Cibulski, S. P., Teixeira, F. & Sales, E. De. A Novel Anelloviridae Species Detected in Tadarida brasiliensis Bats : First Sequence of a Chiropteran Anellovirus. *genome Announ.* **2**, e01028-14 (2014).

13. Hrazdilová, K. *et al.* New species of Torque Teno miniviruses infecting gorillas. *Virology* **487**, 207–214 (2016).

14. Ng, T. *et al.* Metagenomic identification of a novel anellovirus in Pacific harbor seal (Phoca vitulina richardsii) lung samples and its detection in samples from multiple years. *J. Gen. Virol.* **92**, 1318–1323 (2011).

15. Shi, C. *et al.* A Metagenomic Survey of Viral Abundance and Diversity in Mosquitoes from Hubei Province. *PLoS One* **10**, e0129845 (2015).

16. Hsiao, K., Wang, L., Lin, C. & Liu, H. New Phylogenetic Groups of Torque Teno Virus

Identified in Eastern Taiwan Indigenes. *PLoS One* **11**, e0149901 (2016).

17. Virgin, H. W., Wherry, E. J. & Ahmed, R. Redefining Chronic Viral Infection. *Cell* **138**, 30–50 (2009).

18. Spandole, S., Berca, L. M. & Miha, G. Human anelloviruses : an update of molecular , epidemiological and clinical aspects. *Arch. Virol.* **160**, 893–908 (2015).

19. Nishizawa, T. *et al.* A novel DNA virus (TTV) associated with elevated transaminase levels in posttransfusion hepatitis of unknown etiology. *Biochem.Biophys.Res.Commu.* **241**, 92–97 (1997).

20. Biagini, P. *et al.* Familly Anelloviridae. in *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses* (eds. King, A. M. Q., Adams, M. J., Cartens, E. B. & Lefkowitz, E. J.) 331–341 (Elsevier Scientific Publ. Co., 2011).

21. Stremlau, M. H. *et al.* Discovery of Novel Rhabdoviruses in the Blood of Healthy Individuals from West Africa. *PLoS Negl. Trop. Dis.* **9**, e0003631 (2015).

22. Kapoor, A. *et al.* Virome analysis of transfusion recipients reveals a novel human virus that shares genomic features with hepaciviruses and pegiviruses. *MBio* **6**, e01466-15 (2015).

23. Popgeorgiev, N. *et al.* Marseillevirus-like virus recovered from blood donated by asymptomatic humans. *J. Infect. Dis.* **208**, 1042–1050 (2013).

24. Sauvage, V. & Eloit, M. Viral metagenomics and blood safety. *Transfus. Clin. Biol.* **23**, 28–38 (2016).

25. De Vlaminck, I. *et al.* Temporal Response of the Human Virome to Immunosuppression and Antiviral Therapy. *Cell* **155**, 1178–1187 (2013).

26. Conceição-Neto, N. *et al.* Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci. Rep.* **5**, 16532 (2015).

27. Kohl, C. *et al.* Protocol for Metagenomic Virus Detection in Clinical Specimens. *Emerg. Infect. Dis.* **21**, 48–57 (2015).

28. Hall, R. J. *et al.* Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *J. Virol. Methods* **195**, 194–204 (2014).

29. Cheval, J. *et al.* Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. *J. Clin. Microbiol.* **49**, 3268–3275 (2011).

30. Asplund, M. *et al.* Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. *Clin. Microbiol. Infect.* **25**, (2019).

31. Colson, P. *et al.* Evidence of the megavirome in humans. *J. Clin. Virol.* **57**, 191–200 (2013).

32. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).

33. Martí, J. M. Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. *PLoS Comput. Biol.* **15**, e1006967 (2019).

34. Castillo, D. J., Rifkin, R. F., Cowan, D. A. & Potgieter, M. The Healthy Human Blood Microbiome: Fact or Fiction? *Front. Cell. Infect. Microbiol.* **9**, 148 (2019).

35. Ninomiya, M. *et al.* Identification and genomic characterization of a novel human torque teno virus of 3.2 kb. *J. Gen. Virol.* **88**, 1939–1944 (2007).

36. Rosseel, T., Pardon, B., De Clercq, K., Ozhelvaci, O. & Van Borm, S. False-positive results in metagenomic virus discovery: A strong case for follow-up diagnosis. *Transbound. Emerg. Dis.* **61**, 293–299 (2014).

37. Naccache, S. N. *et al.* The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns. *J. Virol.* **87**, 11966–11977 (2013).

38. Fahsbender, E. *et al.* Diverse and highly recombinant anelloviruses associated with Weddell seals in Antarctica. *Virus Evol.* **3**, vex017 (2017).

39. Simmonds, P. *et al.* Consensus statement: Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* **15**, 161–168 (2017).

40. Focosi, D., Antonelli, G., Pistello, M. & Maggi, F. Torquetenovirus : the human virome from bench to bedside. *Clin. Microbiol. Infect.* **22**, 589–593 (2016).

41. Thom, K. & Petrik, J. Progression Towards AIDS Leads to Increased Torque Teno Virus and Torque Teno Minivirus Titers in Tissues of HIV Infected Individuals. *J. Med. Virol.* **79**, 1–7 (2007).

42. Li, L. *et al.* AIDS Alters the Commensal Plasma Virome. *J. Virol.* **87**, 10912–10915 (2013).

43. Spandole-Dinu, S. *et al.* Prevalence of human anelloviruses in Romanian healthy subjects and patients with common pathologies. *BMC Infect. Dis.* **18**, 334 (2018).

44. Béland, K. *et al.* Torque Teno Virus Load as a Biomarker of Immunosuppression? New Hopes and Insights. **210**, 667–70 (2014).

45. Focosi, D., Macera, L., Pistello, M. & Maggi, F. Torque Teno Virus Viremia Correlates With Intensity of Maintenance Immunosuppression in Adult Orthotopic Liver Transplant. *J. Infect. Dis.* **210**, 667–668 (2014).

46. Charest, A. J. *et al.* Global occurrence of Torque teno virus in water systems. *J. Water Health* **13**, 777–789 (2015).

47. Pan, S. *et al.* Identification of a Torque Teno Mini Virus (TTMV) in Hodgkin's Lymphoma Patients. *Front. Microbiol.* **9**, 1680 (2018).

48. Eibach, D. *et al.* Viral metagenomics revealed novel betatorquevirus species in pediatric inpatients with encephalitis / meningoencephalitis from Ghana. *Sci. Rep.* **9**, 2360 (2019).

49. Ng, T. F. F., Dill, J. A., Camus, A. C., Delwart, E. & Meir, E. G. Van. Two new species of betatorqueviruses identified in a human melanoma that metastasized to the brain. *Oncotarget* **8**, 105800–105808 (2017).

50. Zhang, Y. *et al.* A novel species of torque teno mini virus (TTMV) in gingival tissue from chronic periodontitis patients. *Sci. Rep.* **6**, 26739 (2016).

51. Burián, Z. *et al.* Detection and follow-up of torque teno midi virus ('"small anelloviruses"') in nasopharyngeal aspirates and three other human body fluids in children. *Arch.Virol.* **156**, 1537–1541 (2011).

52. Maggi, F. *et al.* Relationships between Total Plasma Load of Torquetenovirus (TTV) and TTV Genogroups Carried. *J. Clin. Microbiol.* **43**, 4807–4810 (2005).

53. Béland, K. *et al.* Torque Teno Virus in Children Who Underwent Orthotopic Liver Transplantation: New Insights About a Common Pathogen. *J. Infect. Dis.* **209**, 247–254 (2014).

54. Gonzales-gustavson, E. *et al.* Identification of sapovirus GV.2, astrovirus VA3 and novel anelloviruses in serum from patients with acute hepatitis of unknown aetiology. *PLoS One* **12**, e0185911 (2017).

55. Burra, P. *et al.* Torque Teno virus: any pathological role in liver transplanted patients? *Transpl. Int.* **21**, 972–979 (2008).

56. Focosi, D. *et al.* Torquetenovirus viremia kinetics after autologous stem cell transplantation are predictable and may serve as a surrogate marker of functional immune reconstitution. *J. Clin. Virol.* **47**, 189–192 (2010).

57. Segura-wang, M., Görzer, I., Jaksch, P. & Puchhammer-stöckl, E. Temporal dynamics of the lung and plasma viromes in lung transplant recipients. *PLoS One* **13**, e0200428 (2018).

58. French, R. K. & Holmes, E. C. An Ecosystems Perspective on Virus Evolution and Emergence. *Trends Microbiol.* **28**, 165–175 (2020).

59. Roossinck, M. J. Plants , viruses and the environment : Ecology and mutualism. *Virology* **479–480**, 271–277 (2015).

60. Kernbauer, E., Ding, Y. & Cadwell, K. An enteric virus can replace the beneficial function of commensal bacteria. *Nature* **516**, 94–98 (2014).

61. Tyschik, E. A., Rasskazova, A. S., Degtyareva, A. V, Rebrikov, D. V & Sukhikh, G. T. Torque teno virus dynamics during the first year of life. *Virol. J.* **15**, 96 (2018).

62. Bédarida, S., Dussol, B., Signoli, M. & Biagini, P. Analysis of Anelloviridae sequences characterized from serial human and animal biological samples. *Infect.Genet.Evol.* **53**, 89–93 (2017).

63. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

64. Bushnell, B., Rood, J. & Singer, E. BBMerge – Accurate paired shotgun read merging via overlap. *PLoS One* **12**, e0185056 (2017).

65. Martí, J. M. & Garay, C. P. Not just BLAST nt: WGS database joins the party. *bioRxiv* 653592 (2019). doi:10.1101/653592

66. Nurk, A. *et al.* Assembling genomes and minimetagenomes from highly chimeric reads. in *Research in Computational Molecular Biology* (eds. Deng, M., Jiang, R., Sun, F. & Zhang, X.) 158–170 (Springer-Verlag Berlin Heidelberg, 2013).

67. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

68. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X : Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).

69. Subramanian, B., Gao, S., Lercher, M. J., Hu, S. & Chen, W. Evolview v3 : a webserver for visualization , annotation , and management of phylogenetic trees. *Nucleic Acids Res.* **47**, 270–275 (2019).

Supplementary Table S1. Centrifuge results of the pilot study. For each extraction (Direct extraction vs. high centrifugation) and filtration (0.45 μM vs. 1.0 μM) treatment, sequencing libraries were prepared independently for DNA and RNA samples.

Supplementary Table S2. Summary of Centrifuge results for the 12 pools analysed. For each pool, the total number of reads passing filtering/trimming analyses, and those classified as human, bacterial, anellovirus and other viruses, are indicated.

Supplementary Table S3. Results of bacterial taxonomic classification using Centrifuge for controls and samples. Recentrifuge values are also provided for samples. Reads are provided at phylum, class and order level.

Supplementary Table S4. Results of viral taxonomic classification using Centrifuge for controls and samples. Recentrifuge values are also provided for samples. The number of total, eukaryotic, human, bacterial and viral reads is shown. Subsequent rows show the distribution of reads in the viral fraction, including spiked viruses, viral families and other taxonomic levels related with unclassified viruses.

Supplementary Table S5. List of sequences/contigs detected in our study with the SPAdes analysis. For each sequence, names in the first column refers to pool (P) and isolate number (c). Contig length (in nucleotides), SPAdes coverage, deduced size of the putative ORF1 (in amino acids), and accession number is given. Last column indicates genus assignment in accordance with phylogenetic and blast analyses. *contigs showing terminal redundancy and, consequently, considered complete genomes. **contigs yielding incomplete ORF1, since nucleotide sequence is interrupted before reaching initiation/stop codon.

Supplementary Table S6. List of anellovirus isolates downloaded from Genbank. Accession number, isolate name and anellovirus genus is indicated. *Those meeting the species demarcation criteria were chosen for subsequent phylogenetic analyses including the sequences described in our study.

Supplementary Table S7. Nucleotide pairwise identity matrix obtained using ORF1 alignment of downloaded TTV representative genotypes and the new viral sequences assigned to this genus.

Supplementary Table S8. Nucleotide pairwise identity matrix obtained using ORF1 alignment of downloaded TTMV representative genotypes and the new viral sequences assigned to this genus.

Supplementary Table S9. Nucleotide pairwise identity matrix obtained using ORF1 alignment of downloaded TTMDV representative genotypes and the new viral sequences assigned to this genus.

Supplementary Table S10. Primers used for PCR amplification and Sanger sequencing of ORF1 gene. Internal primers (F2/R2) were also used for sequencing when necessary.

Supplementary Figure S1. Global phylogenetic tree for the ORF1 of the three anellovirus genera. All downloaded sequences and the sequences described in this study (labelled in red) are included. 0.7-0.85 and 0.85-1.0 bootstrap value ranges are indicated with yellow and red circles, respectively. Scale bar indicates evolutionary distance in nucleotide substitutions per site.

Supplementary Figure S2. Phylogenetic tree including the representative genotypes from TTV genus. The sequences described in this study that can be considered as new species are also included and labelled in red. Numbers between brackets indicate the number of new sequences clustering with a specific genotype (All new sequences are explicitly shown in Fig. 3). Bootstrap values are indicated in red at each branch point. Scale bar indicates evolutionary distance in nucleotide substitutions per site.

Figures

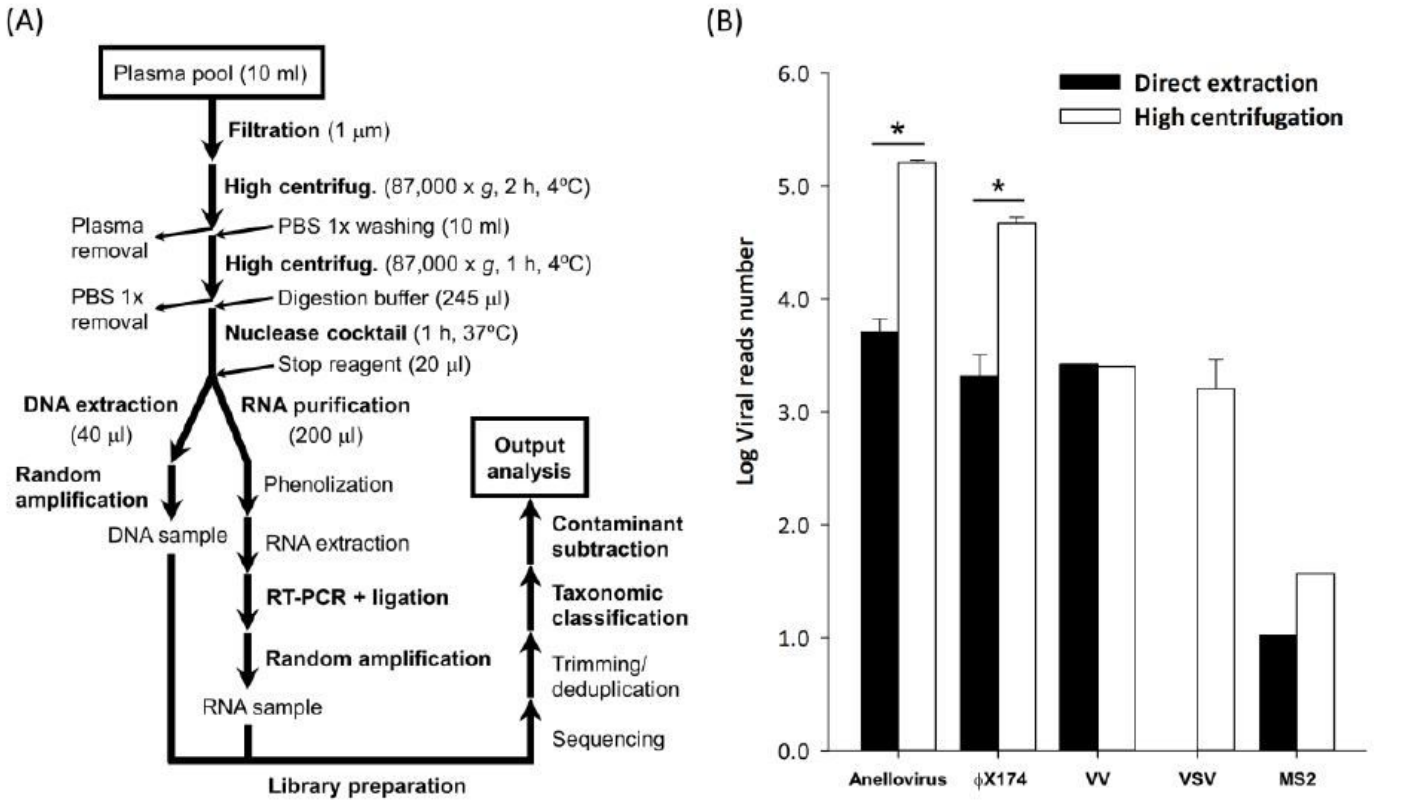


Figure 1

Experimental and bioinformatics workflow (A) and comparison between viral reads recovered with direct extraction from plasma and the protocol involving initial high centrifugation (B). Main steps at panel A are marked in bold (See details in Methods section). For panel B, error bars indicate standard error of the mean (SEM, n = 2 replicates). Asterisk indicates the statistical significance of a log-scale t-test analyzing the efficiency of the purification protocols (*□<0.001). Viral reads per million of total reads are used. For VV, the only indicated value for each treatment was obtained with 1 □m pore size filtration.

(A)

	%reads		
	Human	Bacterial	Viral
Centrifuge	4.4	37.9	57.7
Recentrifuge	0.0	9.2	90.8

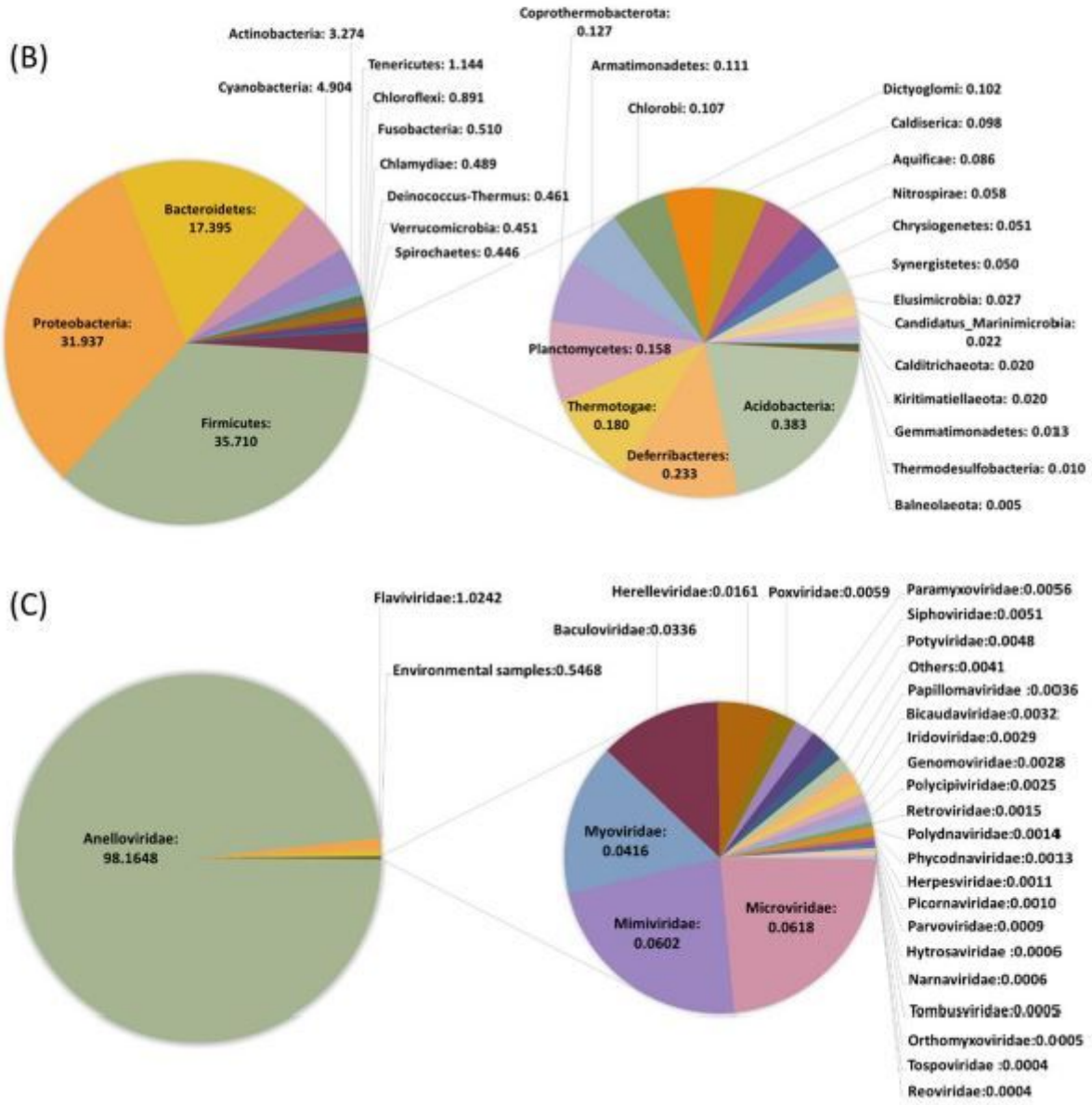


Figure 2

Summary of bioinformatics subtraction for human, bacterial and viral groups (A) and description of the microbiome (B) and the virome (C) characterized in this study. Classification is shown for bacteria and viruses at phylum and family level, respectively. Frequencies were obtained excluding spiked virus contribution in A and C panels.

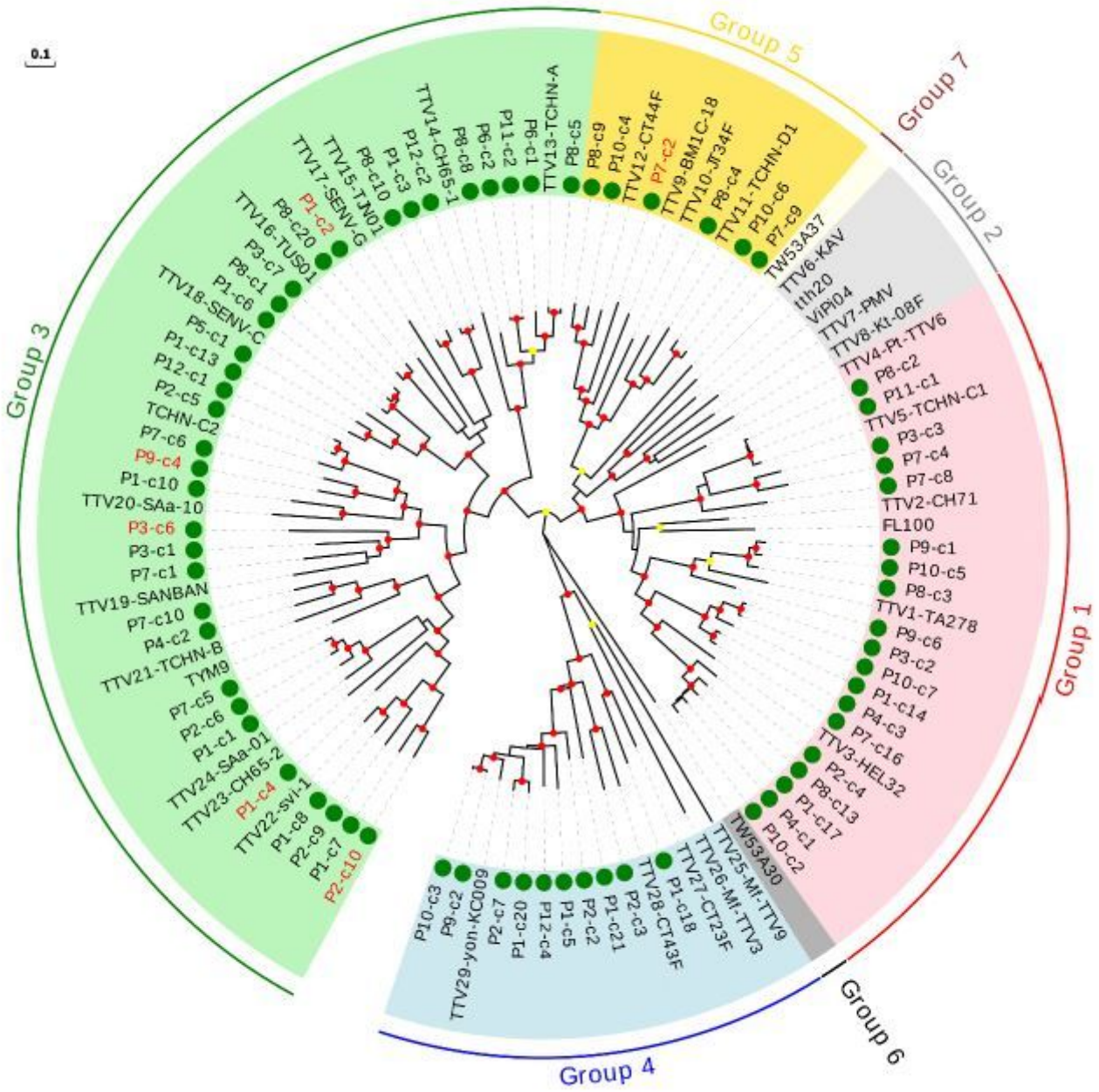


Figure 3

Phylogenetic tree for the ORF1 including the representative genotypes from TTV genus. Sequences described in this study are marked with a green circle. Those sequences that could be considered as new species are labelled in red. 0.7-0.85 and 0.85-1.0 bootstrap value ranges are indicated with yellow and red circles, respectively. Scale bar indicates evolutionary distance in nucleotide substitutions per site.

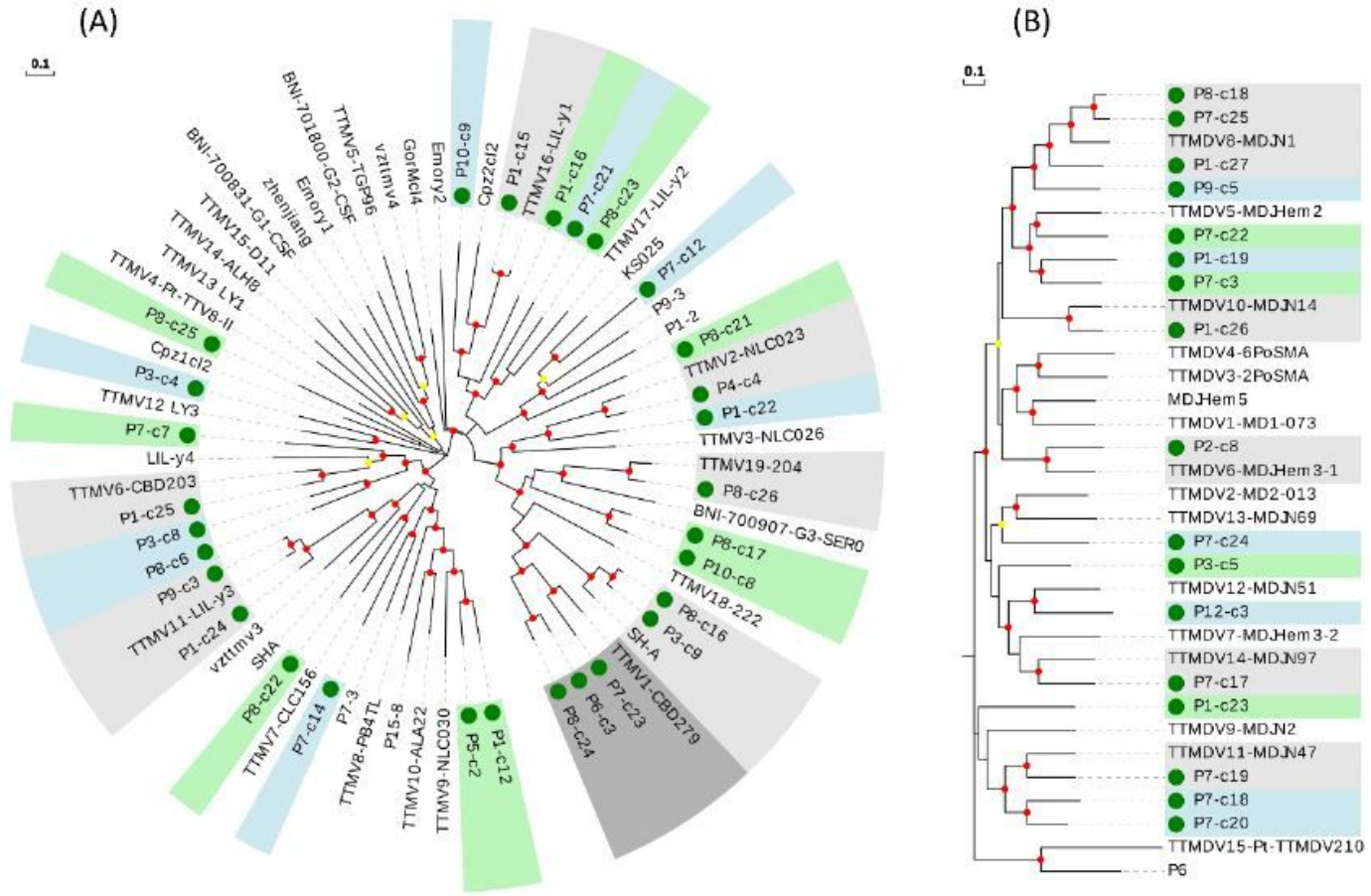


Figure 4

Phylogenetic trees for the ORF1 including the representative genotypes from TTMV (A) and TTMDV (B) genera. Sequences described in this study are marked with a green circle. New species (including one or more new sequences) are indicated with background green or blue color in order to distinguish contiguous clusters. Clusters of representative species including new sequences are indicated with background light or dark grey colors in order to distinguish contiguous clusters. 0.7-0.85 and 0.85-1.0 bootstrap value ranges are indicated with yellow and red circles, respectively. Scale bar indicates evolutionary distance in nucleotide substitutions per site.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS3.xlsx
- SupplementaryTableS1.docx
- SupplementaryTableS4.xlsx
- SupplementaryTableS5.xlsx
- SupplementaryTableS2.docx

- SupplementaryTableS6.xlsx
- SupplementaryTableS3.xlsx
- SupplementaryTableS7.xlsx
- SupplementaryTableS4.xlsx
- SupplementaryTableS8.xlsx
- SupplementaryTableS5.xlsx
- SupplementaryTableS6.xlsx
- SupplementaryTableS9.xlsx
- SupplementaryFigureS1.tiff
- SupplementaryTableS7.xlsx
- SupplementaryTableS8.xlsx
- SupplementaryTableS9.xlsx
- SupplementaryTableS10.docx
- SupplementaryFigureS1.tiff
- SupplementaryFigrueS2.tiff