scientific reports

OPEN

Discovery and comparative genomic analysis of a novel equine anellovirus, representing the first complete *Mutorquevirus* genome

Mathew Fisher[1], Michelle Nebroski[1], Jennifer Davies[2], Eugene Janzen[2], Daniel Sullivan[1] & Oliver Lung[1,3]✉

The complete genome of a novel torque teno virus species (*Torque teno equus virus 2* (TTEqV2) isolate Alberta/2018) was obtained by high-throughput sequencing (HTS) of nucleic acid extracted from the lung and liver tissue of a Quarter Horse gelding that died of nonsuppurative encephalitis in Alberta, Canada. The 2805 nucleotide circular genome is the first complete genome from the *Mutorquevirus* genus and has been approved as a new species by the International Committee on Taxonomy of Viruses. The genome contains several characteristic features of torque teno virus (TTV) genomes, including an ORF1 encoding a putative 631 aa capsid protein with an arginine-rich N-terminus, several rolling circle replication associated amino acid motifs, and a downstream polyadenylation signal. A smaller overlapping ORF2 encodes a protein with an amino acid motif ($WX_7HX_3CXCX_5H$) which, in general, is highly conserved in TTVs and anelloviruses. The UTR contains two GC-rich tracts, two highly conserved 15 nucleotide sequences, and what appears to be an atypical TATA-box sequence also observed in two other TTV genera. Codon usage analysis of TTEqV2 and 11 other selected anelloviruses from five host species revealed a bias toward adenine ending (A3) codons in the anelloviruses, while in contrast, A3 codons were observed at a low frequency in horse and the four other associated host species examined. Phylogenetic analysis of TTV ORF1 sequences available to date shows TTEqV2 clusters with the only other currently reported member of the *Mutorquevirus* genus, Torque teno equus virus 1 (TTEqV1, KR902501). Genome-wide pairwise alignment of TTEqV2 and TTEqV1 shows the absence of several highly conserved TTV features within the UTR of TTEqV1, suggesting it is incomplete and TTEqV2 is the first complete genome within the genus *Mutorquevirus*.

Torque teno viruses (TTVs) (family *Anelloviridae)* are non-enveloped viruses with small, circular, negative-sense, single-stranded DNA genomes that vary in length from 2.1 to 3.9 kb[1]. Anelloviruses are prevalent globally and have been reported in humans as well as a wide variety of wild and domestic animals, including non-human primates (e.g., chimpanzees, macaques, tamarin monkeys, and douroucouli), wild boars, badgers, pine martens, tupaias, rodents, bats, sea turtles, sea lions, livestock (e.g., pigs, sheep, cattle, camels, and poultry) and companion animals (e.g., cats and dogs)[1]. Several diseases have been proposed to be linked with TTV infection; however, there are few reports that support its potential as an etiological agent[1]. The International Committee on Taxonomy of Viruses (ICTV) currently recognizes 30 genera within *Anelloviridae.* Taxonomic classification of anelloviruses is based on nucleotide sequence similarity of ORF1 with cut-off values of 44% and 65%, respectively, for genus and species[2].

TTV genomes reported to date consist of an untranslated region (UTR), two main open reading frames (ORFs), and may also have a variable number of additional ORFs. The UTR contains several conserved genomic features, including at least one GC-rich tract[3] and several transcription elements[4,5]. Additionally, there are two 15 nucleotide conserved sequences (CGAATGGCTGAGTTT and AGGGGCAATTCGGGC) in the UTR of TTVs from both human and animal hosts[5–8]. ORF1 encodes a product of approximately 700–770 amino acids, which is considered the viral capsid protein[7]. Conserved amino acid motifs associated with rolling circle replication (Motif I: Fu[t/u][l/y][t/p], Motif II: [p/u]HuH and Motif III: YxxK) and helicase activity (Walker-A: GxxxxGK[S/T],

[1]National Centre for Foreign Animal Disease, Canadian Food Inspection Agency, Winnipeg, MB, Canada. [2]Faculty of Veterinary Medicine, University of Calgary, Calgary, AB, Canada. [3]Department of Biological Sciences, University of Manitoba, Winnipeg, MB, Canada. ✉email: oliver.lung@inspection.gc.ca

Walker-B: hhxh[D/E][D/E] and motif C: h[T/S/x][T/S/x]N) observed in other single-stranded circular DNA viruses are also present in the ORF1 of some TTVs; however, there is no consistent trend in their presence[9]. ORF2 codes for a product of about 200 amino acids which contains a protein-tyrosine phosphatase amino acid motif ($WX_7HX_3CXCX_5H$) found in both TTVs and chicken anemia virus (a gyrovirus)[10,11] and is thought to be involved in cellular and/or viral protein regulation and processing during natural infection[5,12].

Currently, there is only one publicly available sequence within the genus *Mutorquevirus*, Torque teno equus virus 1 (TTEqV1, KR902501); however, it is missing several features highly conserved within the UTR of TTV genomes. These missing features include one of the two highly conserved 15 nt sequences and a GC-rich region, suggesting that the UTR of the sequence currently available for TTEqV1 is incomplete. Here, we report the complete genome sequence of a novel TTV species, Torque teno equus virus 2 (TTEqV2) isolate Alberta/2018, discovered via high-throughput sequencing (HTS) of tissue samples collected from a horse that died of nonsuppurative encephalitis. The novel virus described here has been officially approved by the ICTV as a novel species within the genus *Mutorquevirus*[13]. The name of the novel virus complies with current ICTV naming guidelines, however switching to a binomial naming system has been suggested by the ICTV but not yet formally adopted[14]. When this system is formally adopted, the name of the novel virus would be changed to Mutorquevirus equid 2.

Materials and methods

Case history.

The carcass of a 2-year-old Quarter Horse gelding from southern Alberta, Canada that died suddenly was submitted to the Diagnostic Services Unit at University of Calgary's Faculty of Veterinary Medicine for post-mortem examination. Gross examination revealed lesions of trauma consistent with the horse being down and thrashing prior to death. Histopathology revealed severe nonsuppurative meningoencephalitis as the cause of death. Immunohistochemistry was negative for rabies virus, West Nile virus and *Sarcocystis neurona*. PCR was negative for eastern equine encephalitis virus and western equine encephalitis virus. Post-mortem liver, lung, spleen, brain and kidney tissue samples were submitted to the Canadian Food Inspection Agency's National Center for Foreign Animal Disease Genomics Unit for characterization via HTS.

Sample processing and high-throughput metagenomic sequencing.

Tissue processing and HTS were performed as previously described[15–17]. Briefly, ten percent suspensions from the liver, lung, spleen, and kidney tissues were processed on a Precellys 24 Dual Tissue Homogenizer (Bertin Instruments). Nucleic acid extraction was performed using the Ambion MagMax Viral RNA Isolation Kit (Thermo Fisher Scientific) according to the manufacturer's instructions and eluted in UltraPure water (Sigma-Aldrich). Brain tissue in formalin solution was processed separately using the Agencourt FormaPure Total kit (Beckman Coulter), designed to extract total nucleic acid from FFPE tissue. Since the brain tissue was not paraffin embedded, the "deparaffinization" step was omitted, but the manufacturer's instructions were followed otherwise. Two extractions were prepared: one from the outside of the brain exposed directly to formalin and the second from the inside of the brain after cutting it in half.

To enable broad metagenomic detection of viruses with either DNA or RNA genomes, reverse transcription was performed separately on extracted nucleic acid from each tissue using the Invitrogen SuperScript IV First-Strand Synthesis System (SSIV) (Thermo Fisher Scientific) according to the manufacturer's instructions using a tagged random nonamer primer (40 µM, GTT TCC CAG TCA CGA TAN NNN NNN NN). Sequenase Version 2.0 DNA Polymerase (Thermo Fisher Scientific) was used to perform second strand synthesis. Sequence-independent single-primer amplification (SISPA) was performed using AccuPrime Taq DNA Polymerase System (Thermo Fisher Scientific) with the manufacturer's recommended conditions. Here, cDNA was amplified using a primer complementary to the tag introduced during reverse transcription. The SISPA product was purified using Genomic DNA Clean & Concentrator-10 columns (Zymo Research), quantified with the Qubit™ dsDNA HS Assay Kit on the Qubit® 3.0 Fluorometer (Thermo Fisher Scientific), and subsequently used for HTS library preparation.

Sequencing library preparation and enrichment were performed individually on each of the tissue derived samples using the Kappa HyperPlus library preparation kit (Roche Diagnostics) and a custom pan-vertebrate virus-targeted enrichment probe panel as previously described[16,18,19]. Following enrichment, pooled libraries were quantified for concentration (Qubit™ dsDNA HS Assay Kit on the Qubit® 3.0 Fluorometer (Thermo Fisher Scientific)) and fragment size (High Sensitivity DNA Kit on the 2100 Bioanalyzer (Agilent Technologies)) and sequenced on the Illumina MiSeq with a V3 flow cell using a 600 cycle kit (Illumina).

Metagenomic sequencing assembly.

Initial exploratory metagenomic analysis was done as previously described[16]. Briefly, an in-house developed automated taxonomic classification workflow (nf-villumina v2.0.0[20]) was used to analyze metagenomic sequencing data. First, nf-villumina removed Illumina PhiX Sequencing Control V3 reads using BBDuk[21], followed by adaptor removal and quality filtering with fastp[22]. Filtered reads were taxonomically classified with Centrifuge[23] and Kraken2[24] using an NCBI nt Centrifuge index built February 14, 2020 and a Kraken2 index of NCBI RefSeq sequences of archaea, bacteria, viral and the human genome GRCh38 downloaded and built on March 22, 2019. Viral and unclassified reads were retained for de novo assembly with Unicycler[25], Shovill[26], and MEGAHIT[27], and the resulting contigs from each were queried against the NCBI nr/nt database (downloaded January 9, 2020) using blastn (v2.9.0)[28] (default parameters except "e-value 1e−6"). Contigs of interest were further analyzed in Geneious (v9.1.8)[29] using a combination of reference assembly with unfiltered reads (default medium–low sensitivity settings and five iterations) and manual alignment-based correction.

Illumina amplicon sequencing.

The partial genome consensus sequence generated from initial metagenomic sequence analysis was used to design primers to generate three PCR amplicons (UTR-Amp1 [Forward:

GAA TGC TCA CAG AGT CTG C, Reverse: TCG GCG TCT TCT CCA]; UTR-Amp2 [Forward: AAG CGA AGG AGA CAT CC, Reverse: TCG GCG TCT TCT CCA]; UTR-Amp3 [Forward: AAG CGA AGG AGA CAT CC, Reverse: AGA ACC TTG CCC AGC]) covering the unsequenced region of the UTR. PCR amplification of the extracted lung and liver-derived nucleic acid was conducted using the SuperScript™ III One-Step RT-PCR System with Platinum™ Taq DNA Polymerase kit (ThermoFisher) according to manufacturer's recommendations. The PCR mixture consisted of 2 µL of extracted nucleic acid, 0.3 µM of each primer, and 2 µL of SuperScript™ III RT/Platinum™ Taq Mix in 1 × reaction buffer in a final volume of 50 µL with UltraPure Distilled Water (Sigma-Aldrich). Amplification conditions were denaturation at 94 °C for 2 min followed by 40 PCR cycles with denaturation at 94 °C for 15 s., annealing at 55 °C for 30 s. and extension at 68 °C for 1 min. with a final extension step of 68 °C for 5 min. PCR product was visualized using a QIAxcel instrument (QIAGEN) and prepared for sequencing using the Nextera XT Library Prep Kit (Illumina), and sequenced on the Illumina MiSeq with a V2 flow cell using a 300 cycle kit (Illumina).

Illumina amplicon sequencing assembly. Amplicon sequencing reads from lung and liver tissue were combined and mapped to the partial genome consensus sequence previously generated from metagenomic sequencing data, using Geneious (v9.1.8)[29] iterative reference assembly (using default medium–low sensitivity settings and five iterations). The consensus sequence was further analyzed using an alignment-based manual correction method.

Nanopore amplicon sequencing. Oxford Nanopore long-read sequencing was used for subsequent amplicon sequencing to generate long reads covering the whole unsequenced portion of the UTR. Previously designed primers were used to generate amplicons UTR-Amp1, UTR-Amp2, and UTR-Amp3 using a PCR enzyme designed for amplification of GC-rich targets. Lung-derived nucleic acid was amplified using Invitrogen Platinum SuperFi II (ThermoFisher) according to manufacturer's recommendations. The PCR mixture consisted of 2 µL of extracted nucleic acid, 0.5 µM of each primer, and 5 uL of 5X SuperFi II Buffer, brought up to a final volume of 20 µL with UltraPure Distilled Water (Sigma-Aldrich). Denaturation was carried out at 98 °C for 30 s. followed by 40 cycles of denaturation at 98 °C for 10 s., annealing at 60 °C for 10 s. and extension at 72 °C for 2 min. with a final extension step of 72 °C for 5 min. The product was visualized using a QIAxcel (QIAGEN) and prepared for sequencing using the Oxford Nanopore Ligation Sequencing Kit (SQK-LSK-109, Oxford Nanopore Technologies) according to manufacturer's recommendations. Sequencing was conducted on a GridION sequencer (Oxford Nanopore Technologies) with live basecalling enabled (high-accuracy basecalling model) using MinKNOW (v20.06.9).

Nanopore amplicon assembly. Nanopore reads were trimmed for adapters using Porechop (v.0.2.4)[30] on default settings. To filter for reads containing the entire region of interest, the trimmed reads were first reference mapped in Geneious (v9.1.8)[29] (default settings, medium sensitivity) to a 40 nt sequence within the amplicon but flanking the unknown region (ATAAAGGCATAGTCCCAATCCCACCAACGCACAAAAAGAG). The resulting mapped reads were then mapped to a 40 nt sequence flanking the opposite side of the unknown region (GAA CGGAGCGAAGCCCGTGGAGTTAAGGGGCAACTCGGGC). The resulting mapped reads containing both known flanking regions were size filtered in Geneious (v9.1.8)[29] to generate a list of reads with a length similar to the estimated amplicon size (1200–1400 nt for Amp1, 950–1350 nt for Amp2 and 725–1125 nt for Amp3). The resulting filtered reads were aligned using MAFFT[31] on default settings, and from this alignment a majority consensus sequence was generated in Geneious (v9.1.8)[29] for each amplicon. The previously unsequenced region of the UTR was extracted from each amplicon, and then they were aligned using MAFFT[31] on default settings. From this alignment, a single consensus sequence was generated in Geneious (v9.1.8)[29]. This consensus sequence representing the unsequenced region of the UTR was added to the previously generated partial genome sequence to generate a preliminary complete genome sequence for further analysis.

Final assembly. Previously generated Illumina amplicon sequencing reads were processed using BBMerge[32], which identified overlapping regions in paired reads, and if present, combined them to generate longer merged reads. Merged reads were mapped to the preliminary complete genome sequence using Geneious (v9.1.8)[29] iterative reference assembly (using default medium–low sensitivity settings and five iterations). The resulting consensus sequence was modified using an alignment-based manual correction method, resulting in a 2,805 nt complete circular genome. As an additional quality check, sequencing reads were mapped to the final consensus sequence using Geneious (v9.1.8)[29] reference assembly (default low sensitivity settings).

Phylogenetic analysis. A maximum-likelihood phylogenetic tree was generated with IQ-TREE[33] from MAFFT[31] multiple sequence alignments (MSA) of the novel genome and representative TTV ORF1 amino acid sequences (n = 146). An IQ-Tree phylogenetic tree was produced using the substitution models indicated in Fig. 1, as selected by ModelFinder[34], with 1000 ultrafast bootstraps[35] and visualized using Interactive Tree Of Life (iTOL)[36]. The resulting tree was pruned to include only the clade containing TTEqV2 and its sister clade (n = 31) as shown in Fig. 1. The whole unpruned tree is shown in Supplementary Material Fig. S1.

Nucleotide composition and codon usage analysis. Nucleotide composition and codon usage analysis were performed using CAIcal[37] (standard genetic code setting) with the same database of representative TTV ORF1 sequences used for phylogenetic analysis (n = 146) with the addition of representative gyrovirus sequences (n = 10). Six TTV sequences were removed from the database and not included in this analysis due to issues that

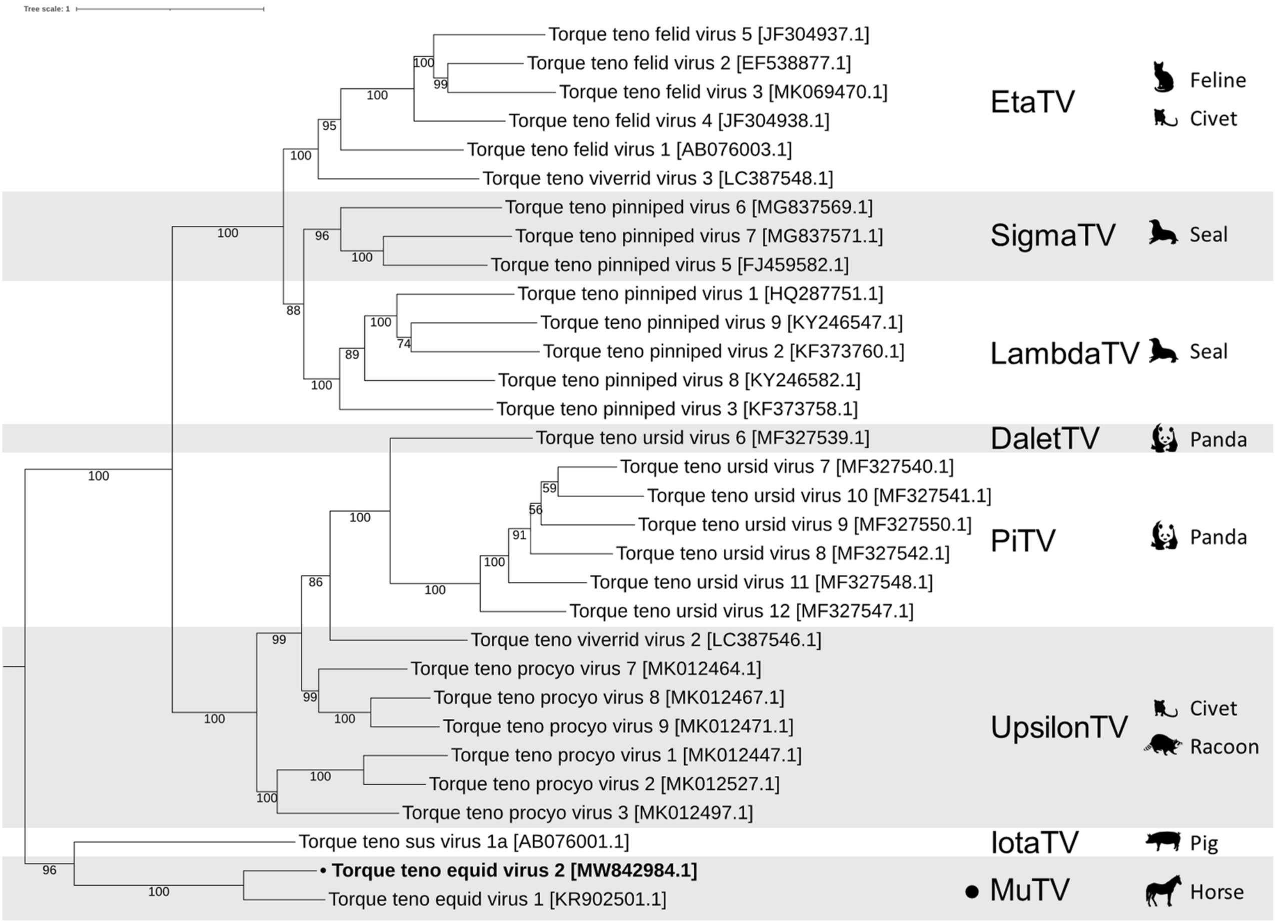
Tree scale: 1
Torque teno felid virus 5 [JF304937.1]
Torque teno felid virus 2 [EF538877.1]
Torque teno felid virus 3 [MK069470.1]
Torque teno felid virus 4 [JF304938.1]
Torque teno felid virus 1 [AB076003.1]
Torque teno viverrid virus 3 [LC387548.1]
Torque teno pinniped virus 6 [MG837569.1]
Torque teno pinniped virus 7 [MG837571.1]
Torque teno pinniped virus 5 [FJ459582.1]
Torque teno pinniped virus 1 [HQ287751.1]
Torque teno pinniped virus 9 [KY246547.1]
Torque teno pinniped virus 2 [KF373760.1]
Torque teno pinniped virus 8 [KY246582.1]
Torque teno pinniped virus 3 [KF373758.1]
Torque teno ursid virus 6 [MF327539.1]
Torque teno ursid virus 7 [MF327540.1]
Torque teno ursid virus 10 [MF327541.1]
Torque teno ursid virus 9 [MF327550.1]
Torque teno ursid virus 8 [MF327542.1]
Torque teno ursid virus 11 [MF327548.1]
Torque teno ursid virus 12 [MF327547.1]
Torque teno viverrid virus 2 [LC387546.1]
Torque teno procyo virus 7 [MK012464.1]
Torque teno procyo virus 8 [MK012467.1]
Torque teno procyo virus 9 [MK012471.1]
Torque teno procyo virus 1 [MK012447.1]
Torque teno procyo virus 2 [MK012527.1]
Torque teno procyo virus 3 [MK012497.1]
Torque teno sus virus 1a [AB076001.1]
• Torque teno equid virus 2 [MW842984.1]
Torque teno equid virus 1 [KR902501.1]
100
99
100
95
100
100
96
100
88
100
74
89
100
100
100
59
56
91
100
100
86
99
99
100
100
100
100
96
100
EtaTV
Feline
Civet
SigmaTV
Seal
LambdaTV
Seal
DaletTV
Panda
PiTV
Panda
UpsilonTV
Civet
Racoon
IotaTV
Pig
• MuTV
Horse

Figure 1. Maximum likelihood phylogenetic trees of representative torque teno viruses. ORF1 amino acid sequences were aligned using MAFFT[31], trees generated using IQ-TREE[33] on find best model setting with ModelFinder[34] (VT + F + R6 model selected) with 1000 ultrafast bootstraps[35] and visualized using iTOL[36]. Genus names are indicated on the right with "-torquevirus" shortened to "TV" and the associated host species for each genus shown. The novel TTEqV2 and its genus (*Mutorquevirus*) and both indicated with a "black filled circle". The tree shown here was pruned to include only the clade containing TTEqV2 and its sister clade with the whole unpruned tree available in Supplementary Material Fig. S1.

caused errors in CAIcal[37]. Several had degenerate bases (AB025946.2, AB038621.1, JF304938.1, KF764701.1, and KX611132.1), while a single sequence (DQ187006.1) had a total number of ORF1 nucleotides not divisible by 3. Relative synonymous codon usage (RSCU) values for host organisms were obtained from a previous study[38]. A spreadsheet containing the nucleotide composition results from CAIcal for all sequences as well as the RSCU values for selected anelloviruses and associated hosts is available in Supplementary Material Table S2.

Results

The workflow used for sequencing and assembly of the complete novel TTV genome utilized a combination of capture probe enrichment, Illumina short-read amplicon sequencing, and Nanopore long-read amplicon sequencing (Fig. 2). Following initial metagenomic sequencing, 2103 and 1892 nt TTV contigs were observed in the lung and liver tissue derived samples, respectively. Nucleic acid derived from other tissues was not incorporated into further analysis because the kidney and spleen samples did not generate TTV contigs while the brain tissue did not generate useful reads, likely due to nucleic acid degradation resulting from storage in formalin. The lung and liver tissue derived contigs were 100% identical in overlapping regions, and were thus combined to generate a single 2267 nt consensus sequence.

Alignment to existing TTV sequences suggested that a portion of the UTR was missing from the 2267 nt consensus sequence. This missing region was determined using amplicon sequencing with primers targeting the regions flanking the missing region, determined by initial metagenomics sequencing. Initial Illumina short-read amplicon sequencing did not generate the complete missing region, likely due to difficulty assembling several GC-rich homopolymeric regions and lack of a suitable reference genome for read mapping. Subsequent Nanopore long-read sequencing was used to generate a scaffold which was used in combination with existing Illumina data to generate a sequence for the entire amplified region. Consensus sequences from metagenomic and amplicon sequencing were combined to generate a 2805 nt final consensus sequence with 50% GC content.

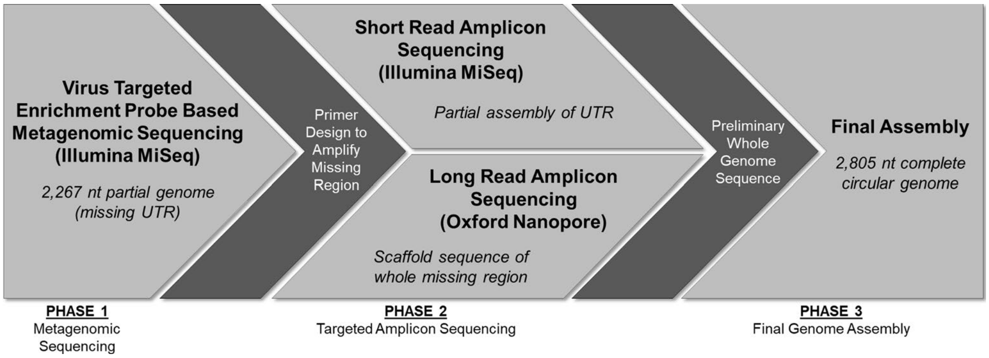
Virus Targeted Enrichment Probe Based Metagenomic Sequencing (Illumina MiSeq)
2,267 nt partial genome (missing UTR)
Primer Design to Amplify Missing Region
Short Read Amplicon Sequencing (Illumina MiSeq)
Partial assembly of UTR
Long Read Amplicon Sequencing (Oxford Nanopore)
Scaffold sequence of whole missing region
Preliminary Whole Genome Sequence
Final Assembly
2,805 nt complete circular genome
PHASE 1
Metagenomic Sequencing
PHASE 2
Targeted Amplicon Sequencing
PHASE 3
Final Genome Assembly

Figure 2. Schematic of workflow used for generation of novel TTEqV2 complete circular genome sequence.

Both metagenomic and amplicon sequencing reads mapped across the entire final consensus sequence, including the linearization point of the circular genome, indicating this sequence represents the complete circular TTEqV2 genome. The structure of the complete circular TTEqV2 genome is shown in Fig. 3.

The TTEqV2 genome shares several common characteristics with previously reported TTVs. ORF1 is the longest ORF and encodes a 631 aa protein with an arginine-rich N-terminus (MAYYWN**R**NNW**RRRR**GAWS**RRR**YYW**RRR**NY**RR**W**RRRRR**V**RR**Q**RRRR**VA**RR**), conserved amino acid motifs and a downstream polyadenylation signal (Fig. 3). While ORF1 encodes a capsid protein, it contains four rolling circle replication (RCR) or helicase activity associated amino acid motifs (two RCR IIIs [YGPK and YLTK], one Walker-A [GTSQQGKT] and one Walker-B [LLTTDE]) that have also been found in other circular ssDNA virus genomes[9]. There is also an ORF2 encoding a 71 aa putative protein, in the same orientation and overlapping with the N-terminal end of ORF1, that contains a highly conserved $WX_7HX_3CXCX_5H$ motif. Based on analysis using SnapGene Viewer v5.0.7 (snapgene.com), the novel TTV genome contains six additional ORFs, encoding hypothetical proteins > 50 aa, ranging in size from 59 to 137 aa. HMMer3 hmmsearch[39] against the Pfam HMM DB (v33.1)[40] (performed February 22, 2023) showed that ORF1 and 2 matched ORF1 and 2 from other TTVs, respectively, while the six additional ORFs had no matches. Blastp[28] analysis on default settings using the nr database showed consistent results, with matches to TTV ORFs for ORF1 and 2 but no matches for any of the six additional ORFs (performed February 22, 2023).

Like other TTVs, the novel genome contains a UTR with two 15 nt conserved motifs, putative transcription factors (TATA box, Sp1 site, Cap site, and polyadenylation signal), and GC-rich region (Figs. 3 and 4). The two 15 nucleotide conserved sequence motifs within the UTR, hereinafter referred to as UTR motif 1 and 2, are similar to those previously described in other TTV genomes[5–7] (Fig. 4). When compared to these previously reported UTR motifs, TTEqV2's UTR motif 1 has 100% identity (CGAATGGCTGAGTTT) while motif 2 has a single nucleotide substitution (AGGGGCAA[T>C]TCGGGC). The TATA box, identified based on its position relative to UTR motif 1 (13 bp upstream), appears to be conserved in TTVs as shown in the alignment in Fig. 4.

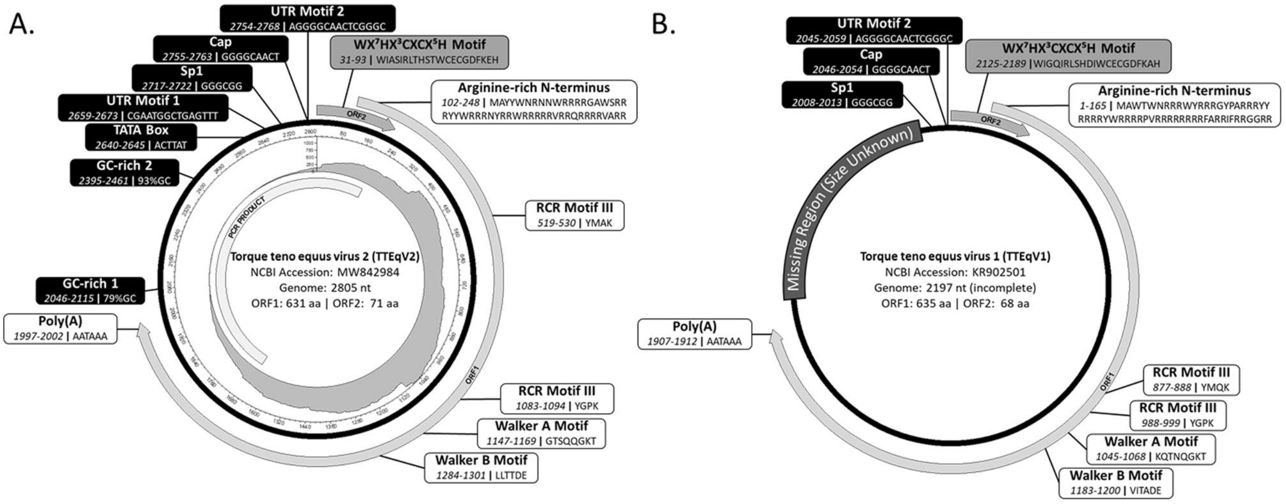

Figure 3. Genomic structure of (**A**) Torque teno equus virus 2 (TTEqV2) and (**B**) Torque teno equus virus 1 (TTEqV1). Inside the annotated circular genome for TTEqV2 is a plot showing the coverage at each nucleotide position (generated using circlize v0.4.15[58]). The region sequenced via subsequent targeted amplicon sequencing is indicated by a bar within the coverage plot. For both viruses, genomic features of interest are labelled with the feature name followed by the position and sequence (or GC content for GC-rich regions in TTEqV2) below.

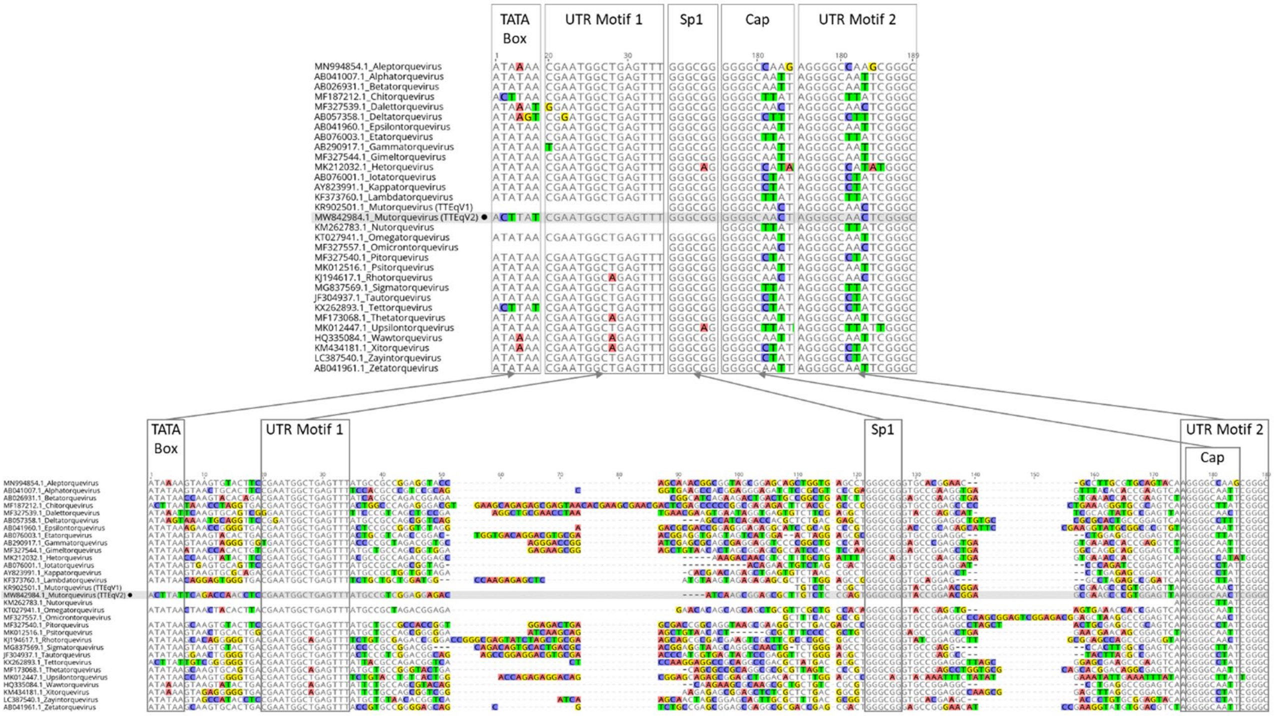
TATA Box
UTR Motif 1
Sp1
Cap
UTR Motif 2
MN994854.1_Aleptorquevirus
AB041007.1_Alphatorquevirus
AB026931.1_Betatorquevirus
MF187212.1_Chitorquevirus
MF327539.1_Dalettorquevirus
AB057358.1_Deltatorquevirus
AB041960.1_Epsilontorquevirus
AB076003.1_Etatorquevirus
AB290917.1_Gammatorquevirus
MF327544.1_Gimeltorquevirus
MK212032.1_Hetorquevirus
AB076001.1_Iotatorquevirus
AY823991.1_Kappatorquevirus
KF373760.1_Lambdatorquevirus
KR902501.1_Mutorquevirus (TTEqV1)
MW842984.1_Mutorquevirus (TTEqV2) ●
KM262783.1_Nutorquevirus
KT027941.1_Omegatorquevirus
MF327557.1_Omicrontorquevirus
MF327540.1_Pitorquevirus
MK012516.1_Psitorquevirus
KJ194617.1_Rhotorquevirus
MG837569.1_Sigmatorquevirus
JF304937.1_Tautorquevirus
KX262893.1_Tettorquevirus
MF173068.1_Thetatorquevirus
MK012447.1_Upsilontorquevirus
HQ335084.1_Wawtorquevirus
KM434181.1_Xitorquevirus
LC387540.1_Zayintorquevirus
AB041961.1_Zetatorquevirus

Figure 4. Nucleotide alignment of conserved region of the UTR from TTV genomes representing every currently recognized genus. Putative transcription factors and conserved motifs are labelled and outlined with boxes. Highlighted bases indicate those that differ from the majority consensus sequence. Each sequence is labelled with the NCBI accession number and genus. The novel TTEqV2 sequence is indicated with a black filled circle and shaded in gray for emphasis.

Interestingly, TTEqV2 contains an atypical putative TATA box (ACTTAT) which differs from the canonical TATA box seen in most TTVs (ATATAA).

The most closely related publically available genome to TTEqV2 is TTEqV1 (KR902501) with an ORF1 pairwise nucleotide identity of 59.7% and amino acid identity of 52.5%. A phylogenetic tree built using representative TTV ORF1 sequences demonstrates that the novel TTEqV2 clusters with TTEqV1 (Fig. 1). Pairwise alignment of the two TTV equine sequences with MAFFT[31] using the default settings shows that TTEqV1 is missing several genomic features conserved in TTVs, including UTR motif 1 and a GC-rich region (Figs. 3 and 4).

Nucleotide composition analysis of ORF1 determined that in both TTEqV2 and TTEqV1, adenine was the most abundant nucleotide at 36.5% and 35.4%, respectively. When similar analysis was performed on a database of representative anellovirus sequences a similar trend was seen, with 138 of 150 (92%) total sequences having adenine as the most abundant nucleotide, an average abundance of 35.5% and a minimum abundance of 24.4%. When gyroviruses were removed, the number of sequences with adenine as the most abundant changed to 134 out of 140 (95.7%), with an average abundance of 35.9%. Interestingly, the six TTV sequences where adenine was not the most abundant all had cytosine as the most abundant nucleotide and came from either a primate (KP296853.1 [27.3%A], KP296854.1 [29.4%A], KP296856.1 [30.5%A] and AB041961.1 [24.4%A]) or feline host (KX262893.1 [28.1%A] and AB076003.1 [26.3%A]) (Supplemental Material S1).

Codon usage analysis of anellovirus ORF1 sequences from genomes representing eight genera, selected based on the availability of codon usage data for associated host species, revealed a bias toward adenine ending (A3) codons in the anelloviruses (Fig. 5). Here, relative synonymous codon usage (RSCU), a measure of the frequency of a specific synonymous codon versus the expected frequency without bias, was used to compare codon usage patterns among anellovirus and host genomes. All anellovirus ORF1 sequences analyzed had at least one overrepresented A3 codon (RSCU > 1.6). Similar analysis performed on the associated host species (horse, swine, canine, human, and chicken)[38] found that none of them had a single overrepresented A3 codon. Analysis of underrepresented codons (RSCU < 0.6) determined that of the twelve total anellovirus ORF1 sequences analyzed, three of the human TTVs had a single underrepresented A3 codon, while none of the six non-human TTVs had any underrepresented A3 codons. It is worth noting that the underrepresented A3 codon for all three human TTVs was CGA. In all cases AGA, another A3 codon which codes for the same amino acid (arginine), was highly overrepresented (all RSCU > 3). The number of under-represented A3 codons ranged from 0 to 6 in the gyroviruses and from 5 to 6 in the analyzed host species. Average RSCU values for A3 codons were greater than or equal to 1 for all anelloviruses and less than 1 for all host species analyzed.

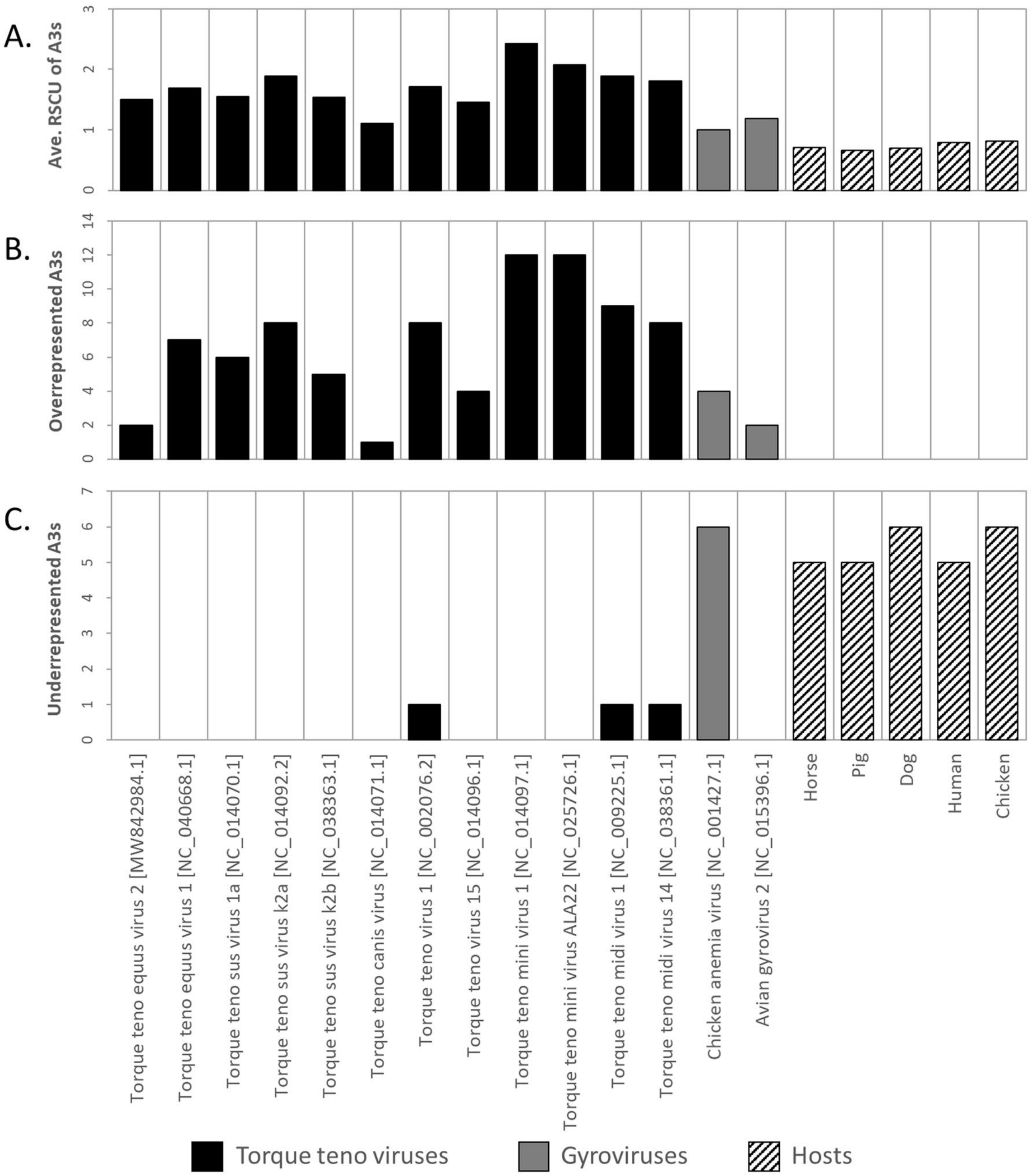

Figure 5. Codon usage analysis of adenine ended codons (A3s) in TTVs, gyroviruses and associated host species. (**A**) Average relative synonymous codon usage (RSCU) of A3s. (**B**) Number of overrepresented A3s (those with RSCU values > 1.6). (**C**) Number of underrepresented A3s (those with RSCU values < 0.6). A spreadsheet containing the RSCU values for the anelloviruses and associated hosts shown here is available in Supplementary Material Table S2. All RSCU values for host organisms were obtained from a previous study[38].

Discussion

A 2,197 nucleotide TTEqV1 genome identified in the metagenomic analysis of plasma from a horse was previously the only sequence within the genus *Mutorquevirus*[41]. Our analysis suggests that the reported TTEqV1 genome sequence is incomplete and missing a portion of the UTR region including one of the two conserved 15 nt sequences and a GC-rich tract, both highly conserved features in the UTR of TTVs. An ORF with homology to the ORF2 identified in TTEqV2 (including the highly conserved $WX_7HX_3CXCX_5H$ motif) was also observed in our analysis of TTEqV1, but is not annotated on the NCBI entry.

The novel TTEqV2 genome contains several genomic features with varying levels of similarity to those previously described in other TTV genomes. ORF1 and ORF2 have similar size, position, and amino acid motifs to other publicly available TTV sequences. TTEqV2 and TTEqV1 have similar amino acid motifs within ORF1; however, some differ in position and/or sequence. The ORF1 of both genomes contain two RCR motif IIIs, one of which is in a similar position and has an identical amino acid sequence (YGPK), while the other has both a different position and sequence (YMQK in TTEqV1, YMAK in TTEqV2). The Walker-A and B motifs are in a similar position in both genomes but differ in amino acid sequence (KQTNQGKT for Walker-A and VITADE for Walker-B in TTEqV1, GTSQQGKT for Walker-A and LLTTDE for Walker-B in TTEqV2).

Two GC-rich regions, characteristic of TTV genomes, are located within the UTR of TTEqV2. The first is 70 nt with 78.6% GC, while the second is 67 nt with 92.5% GC. These GC-rich regions, which contain long homopolymeric stretches, were likely the reason initial analysis with only metagenomic data failed to generate a complete genome sequence. Assembly of the final genome required a combination of metagenomic, short, and long-read amplicon sequencing. Similarly, when the first human TTVs were sequenced, it was thought to be a linear genome due to difficulty amplifying and sequencing GC-rich regions[42].

Transcription regulatory sites identified in TTEqV2, including the Sp1 site, cap site, and polyadenylation signal, are similar to those characterized in other TTV genomes. The Sp1 site and polyadenylation signal exactly match those described in previously characterized TTV genomes, while the Cap site has a single nucleotide difference which is also seen in TTEqV1 (GGGGCAA[T>C]T)[4,5]. The TATA-box, which is well conserved in most TTV genomes, appears to be either heavily modified or missing from the expected region of TTEqV2. Generally, TTV genomes have a TATA-box that is 13 nt upstream of UTR motif 1 and conforms to the canonical consensus sequence (ATATAA) with slight variations in some cases. The putative atypical TATA-box in TTEqV2 (ACT TAT), determined based on location relative to the conserved motif, has three nucleotide differences compared to the canonical sequence. The incomplete TTEqV1 genome does not include UTR motif 1 or the upstream region containing the TATA-box, so the sequence of this region in the other available *Mutorqevirus* genomes is unknown. However, an identical atypical putative TATA-box is seen in the representative *Tettorquevirus* genome (KX262893.1), and one with a single base difference (ACTTA**A**) is seen in the representative *Chitorquevirus* genome (MF187212.1). Both of these representative genomes are the only publically available species within their genus, so whether this atypical putative TATA box is conserved in other sequences of the genus is unknown. Interestingly, neither of these sequences cluster with TTEqV2 based on the alignment of ORF1 and come from different host species (*Tettorquevirus* from feline and *Chitorquevirus* from lemur).

Nucleotide composition analysis revealed that anellovirus ORF1 sequences tend to be adenine rich, with A3 codons favoured in the sequences analyzed. Previous studies made similar observations in anelloviruses[43], swine TTV[44] and equine influenza virus sequences[38]. Interestingly, the opposite trend was observed in the associated host species (horse, pig, dog, human and chicken) for all anellovirus genera analyzed, where A3 codons were underrepresented. A previous study suggested that if codon usage bias in a virus is too similar to that of the host, host translation may be impeded, leading to a greater chance of the virus generating a symptomatic response in the host[45]. The significance of the observation that the TTEqV2 genome has dissimilar codon usage compared to its equine host remains to be determined.

Although TTV has been proposed to be related to many diseases, there are only a few reports supporting the disease-inducing potential of TTV[1]. Human TTVs have been proposed to play a role in the pathogenesis of certain diseases, such as hepatitis[46], hematological disorders[47], respiratory diseases[48], rheumatic autoimmune disease[49]. A recent viral metagenomic study identified a novel betatorquevirus species prevalent in pediatric encephalitis/meningoencephalitis cases, but absent in healthy cohorts[5].

Torque teno sus viruses (TTSuVs) have been found at a particularly high frequency in healthy swine[50,51]. While considered non-pathogenic on their own, there is increasing evidence that TTSuVs may influence the development or outcome of some diseases[52]. For example, co-infection with porcine circovirus type 2 (PCV2) and the associated porcine circovirus diseases deserve special attention[53]. TTSuVs have also been partially attributed to inducing porcine reproductive and respiratory syndrome, porcine dermatitis and nephropathy syndrome, and hepatitis[54,55]. TTSuV2 viremia may be associated with the level of immunocompetence of the animals[52]. A study with pigs infected with hepatitis E virus has shown a correlation between TTSuV and the increased risk of developing severe hepatitis in animals co-infected with PCV2[56]. A high prevalence of TTSuV1, but not TTSuV2, in pigs suffering from porcine respiratory disease complex has been shown[57]. Such viruses would likely be considered components of the host microbiota and unable to cause disease directly, but instead available to be engaged in physiological processes and modulate the organism's response to other pathogens[1]. The relationship between TTV, disease and host immune response is not well understood and therefore the connection between TTEqV2 and the disease observed in the horse, if any even exists, remains to be determined.

In conclusion, this study describes the discovery of a novel anellovirus species which represents the first complete genome within the genus *Mutorquevirus*. Comparative genomic analysis showed that TTEqV2 shares many conserved features with previously reported TTVs and it has been recognized as a novel species by the ICTV[13]. This, along with previous studies using similar methods[15,16,18,19] demonstrates the power of HTS for characterization of unexpected and/or novel viruses in a variety of hosts and sample types.

Data availability

The complete genome is available on NCBI under accession MW842984.

Received: 30 September 2022; Accepted: 2 March 2023
Published online: 06 March 2023

References

1. Manzin, A., Mallus, F., Macera, L., Maggi, F. & Blois, S. Global impact of Torque teno virus infection in wild and domesticated animals. *J. Infect. Dev. Countries* **9**, 562–570 (2015).
2. Biagini, P. *et al.* Family Anelloviridae. In *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*, 331–341 (2011).

3. Kamada, K., Kamahora, T., Kabat, P. & Hino, S. Transcriptional regulation of TT virus: Promoter and enhancer regions in the 1.2-kb noncoding region. *Virology* **321**, 341–348 (2004).
4. Kapusinszky, B. *et al.* Local virus extinctions following a host population bottleneck. *J. Virol.* **89**, 8152–8161 (2015).
5. Eibach, D. *et al.* Viral metagenomics revealed novel betatorquevirus species in pediatric inpatients with encephalitis/meningoencephalitis from Ghana. *Sci. Rep.* **9**, 1–10 (2019).
6. Okamoto, H. *et al.* Species-specific TT viruses in humans and nonhuman primates and their phylogenetic relatedness. *Virology* **277**, 368–378 (2000).
7. Okamoto, H. *et al.* Genomic characterization of TT viruses (TTVs) in pigs, cats and dogs and their relatedness with species-specific TTVs in primates and tupaias. *J. Gen. Virol.* **83**, 1291–1297 (2002).
8. Hu, Y.-W. *et al.* Molecular detection method for all known genotypes of TT Virus (TTV) and TTV-like viruses in thalassemia patients and healthy individuals. *J. Clin. Microbiol.* **43**, 3747–3754 (2005).
9. Rosario, K., Duffy, S. & Breitbart, M. A field guide to eukaryotic circular single-stranded DNA viruses: Insights gained from metagenomics. *Arch. Virol.* **157**, 1851–1871 (2012).
10. Martínez-Guinó, L., Ballester, M., Segalés, J. & Kekarainen, T. Expression profile and subcellular localization of Torque teno sus virus proteins. *J. Gen. Virol.* **92**, 2446–2457 (2011).
11. Peters, M. A., Jackson, D. C., Crabb, B. S. & Browning, G. F. Chicken anemia virus VP2 is a novel dual specificity protein phosphatase. *J. Biol. Chem.* **277**, 39566–39573 (2002).
12. Zheng, H. *et al.* Torque teno virus (SANBAN isolate) ORF2 protein suppresses NF-kappaB pathways via interaction with IkappaB kinases. *J. Virol.* **81**, 11917–11924 (2007).
13. Walker, P. J. *et al.* Recent changes to virus taxonomy ratified by the international committee on taxonomy of viruses (2022). *Arch. Virol.* https://doi.org/10.1007/s00705-022-05516-5 (2022).
14. Walker, P. J. *et al.* Changes to virus taxonomy and to the international code of virus classification and nomenclature ratified by the international committee on taxonomy of viruses (2021). *Arch. Virol.* **166**, 2633–2648 (2021).
15. Papineau, A. *et al.* Genome organization of Canada goose coronavirus, a novel species identified in a mass die-off of Canada geese. *Sci. Rep.* **9**, 5954 (2019).
16. Fisher, M. *et al.* Discovery and comparative genomic analysis of elk circovirus (ElkCV), a novel circovirus species and the first reported from a cervid host. *Sci. Rep.* **10**, 19548 (2020).
17. Lung, O. *et al.* First whole-genome sequence of Cervid atadenovirus A outside of the United States from an Adenoviral hemorrhagic disease epizootic of black-tailed deer in Canada. *Sci. Rep.* https://doi.org/10.1101/2022.02.10.479879 (2022).
18. Wylie, T. N., Wylie, K. M., Herter, B. N. & Storch, G. A. Enhanced virome sequencing using targeted sequence capture. *Genome Res.* **25**, 1910–1920 (2015).
19. Lung, O. *et al. Comparative Genomics Analysis between frog Virus 3-like Ranavirus from the First Canadian Reptile Mortality Event and Similar Viruses from Amphibians*. https://www.researchsquare.com/article/rs-943897/v1 (2021). https://doi.org/10.21203/rs.3.rs-943897/v1.
20. Kruczkiewicz, P. *peterk87/nf-villumina*. https://github.com/peterk87/nf-villumina (2020).
21. Bushnell, B. *BBMap*. https://sourceforge.net/projects/bbmap/.
22. fastp: An Ultra-Fast All-in-One FASTQ Preprocessor | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/34/17/i884/5093234.
23. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *genome res.* **26**, 1721–1729 (2016).
24. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
25. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, 1–10 (2017).
26. Seemann, T. *tseemann/shovill*. https://github.com/tseemann/shovill (2020).
27. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
28. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
29. Kearse, M. *et al.* Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
30. Wick, R. *rrwick/Porechop*. https://github.com/rrwick/Porechop (2020).
31. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
32. Bushnell, B., Rood, J. & Singer, E. BBMerge: Accurate paired shotgun read merging via overlap. *PLoS ONE* **12**, e0185056 (2017).
33. Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A. & Minh, B. Q. W-IQ-TREE: A fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* **44**, W232–W235 (2016).
34. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
35. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
36. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **47**, W256 (2019).
37. Puigbò, P., Bravo, I. G. & Garcia-Vallve, S. CAIcal: A combined set of tools to assess codon usage adaptation. *Biol. Direct* **3**, 38 (2008).
38. Kumar, N. *et al.* Revelation of influencing factors in overall codon usage bias of equine influenza viruses. *PLoS ONE* **11**, e0154376 (2016).
39. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
40. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
41. Li, L. *et al.* Exploring the virome of diseased horses. *J. Gen. Virol.* **96**, 2721–2733 (2015).
42. Bendinelli, M. *et al.* Molecular properties, biology, and clinical implications of TT virus, a recently identified widespread infectious agent of humans. *Clin. Microbiol. Rev.* **14**, 98–113 (2001).
43. Deb, B., Uddin, A. & Chakraborty, S. Composition, codon usage pattern, protein properties, and influencing factors in the genomes of members of the family Anelloviridae. *Arch. Virol.* **166**, 461–474 (2021).
44. Li, G. *et al.* Genetic analysis and evolutionary changes of the torque teno sus virus. *Int. J. Mol. Sci.* **20**, E2881 (2019).
45. Chen, F. *et al.* Dissimilation of synonymous codon usage bias in virus-host coevolution due to translational selection. *Nat. Ecol. Evol.* **4**, 589–600 (2020).
46. Asim, M., Singla, R., Gupta, R. K. & Kar, P. Clinical & molecular characterization of human TT virus in different liver diseases. *Indian J. Med. Res.* **131**, 545–554 (2010).
47. 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).
48. Maggi, F. *et al.* TT virus in the nasal secretions of children with acute respiratory diseases: Relations to viremia and disease severity. *J. Virol.* **77**, 2418–2425 (2003).

49. Gergely, P., Perl, A. & Poór, G. Possible pathogenic nature of the recently discovered TT virus: Does it play a role in autoimmune rheumatic diseases?. *Autoimmun. Rev.* **6**, 5–9 (2006).
50. Blois, S. *et al.* High prevalence of co-infection with multiple Torque teno sus virus species in Italian pig herds. *PLoS ONE* **9**, e113720 (2014).
51. Sibila, M. *et al.* Swine torque teno virus (TTV) infection and excretion dynamics in conventional pig farms. *Vet. Microbiol.* **139**, 213–218 (2009).
52. Kekarainen, T., Sibila, M. & Segalés, J. Prevalence of swine Torque teno virus in post-weaning multisystemic wasting syndrome (PMWS)-affected and non-PMWS-affected pigs in Spain. *J. Gen. Virol.* **87**, 833–837 (2006).
53. Kekarainen, T. & Segalés, J. Torque teno sus virus in pigs: An emerging pathogen?. *Transbound Emerg. Dis.* **59**(Suppl 1), 103–108 (2012).
54. Krakowka, S. *et al.* Evaluation of induction of porcine dermatitis and nephropathy syndrome in gnotobiotic pigs with negative results for porcine circovirus type 2. *Am. J. Vet. Res.* **69**, 1615–1622 (2008).
55. Aramouni, M. *et al.* Torque teno sus virus 1 and 2 viral loads in postweaning multisystemic wasting syndrome (PMWS) and porcine dermatitis and nephropathy syndrome (PDNS) affected pigs. *Vet. Microbiol.* **153**, 377–381 (2011).
56. Savic, B. *et al.* Detection rates of the swine torque teno viruses (TTVs), porcine circovirus type 2 (PCV2) and hepatitis E virus (HEV) in the livers of pigs with hepatitis. *Vet. Res. Commun.* **34**, 641–648 (2010).
57. Rammohan, L. *et al.* Increased prevalence of torque teno viruses in porcine respiratory disease complex affected pigs. *Vet. Microbiol.* **157**, 61–68 (2012).
58. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. Circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).

Acknowledgements

The authors acknowledge partial funding from Canadian Food Inspection Agency (CFIA) Project WIN-A-1408 and Canadian Safety and Security Program Project TI-2222. The authors would also like to acknowledge Dr. Oksana Vernygora and Josip Rudar for feedback on the manuscript and Thomas Harrison for technical assistance.

Author contributions

O.L. and J.D. conceptualized the project. M.F. and O.L. performed experimental design, analyzed the data, and wrote the manuscript's text. M.F. performed experimental work, performed bioinformatics analysis, and made Figs. 1, 2, 3, 4 and 5. M.N. performed sequence alignment and phylogenetic analysis and generated the raw phylogenetic tree used in Fig. 1. D.S. generated the circular coverage map used in Fig. 3. J.D. and E.J performed the necropsy, sample collection, and histopathology. All authors contributed to and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-30875-7.

Correspondence and requests for materials should be addressed to O.L.