Metagenomic identification of viral sequences in laboratory reagents

1.1 Author names

Ashleigh F. Porter[1], Joanna Cobbin[2], Cixiu Li[3], John-Sebastian Eden[2,4], Edward C. Holmes[2]*

1.2 Affiliations

[1] The Peter Doherty Institute of Immunity and Infection, Department of Microbiology and Immunity, University of Melbourne, Melbourne, VIC 3000, Australia.

[2] Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia.

[3] Key Laboratory of Etiology and Epidemiology of Emerging Infectious Diseases in Universities of Shandong, Shandong First Medical University & Shandong Academy of Medical Sciences, Taian 271000, China.

[4] Centre for Virus Research, Westmead Institute for Medical Research, Westmead, Australia.

1.3 Corresponding author

Edward C. Holmes, edward.holmes@sydney.edu.au

1.4 Keywords

Reagent contamination, virology, metatranscriptomics, *Circoviridae, Totiviridae*, *Tombusviridae*, *Lentiviridae*

1.5 Repositories

The viral genome sequence data generated in this study has been deposited in the NCBI database under accession numbers MZ824225-MZ824237. Sequence reads are available at the public Sequence Read Archive (SRA) database with accession SRX6803604 and under the BioProject accession PRJNA735051 reference numbers SRR14737466-71 and BioSample numbers SAMN20355437-40.

2. Abstract

Metagenomic next-generation sequencing has transformed the discovery and diagnosis of infectious disease, with the power to characterize the complete 'infectome' (bacteria, viruses, fungi, parasites) of an individual host organism. However, the identification of novel pathogens has been complicated by widespread microbial contamination in commonly used laboratory reagents. Using total RNA sequencing ("metatranscriptomics") we documented the presence of contaminant viral sequences in multiple libraries of 'blank' negative control sequencing libraries that comprise a sterile water and reagent mix. Accordingly, we identified 14 viral sequences in 7 negative control sequencing libraries. As in previous studies, several circular replication-associated protein encoding (CRESS) DNA virus-like sequences were recovered in the blank libraries, as well as contaminating sequences from the RNA virus families *Totiviridae*, *Tombusviridae* and *Lentiviridae*. These data suggest that the contamination of common laboratory reagents is likely widespread and can comprise a wide variety of viruses.

3. Data summary

The authors confirm all supporting data, code and protocols have been provided within the article or through supplementary data files.

4. Introduction

Culture-independent methods, particularly metagenomic next-generation sequencing (mNGS), have revolutionised pathogen discovery, streamlined pathways of clinical diagnosis, and have enhanced our ability to track infectious disease outbreaks [1], including the current COVID-19 pandemic [2, 3]. These methods can reveal the complete profile of pathogenic and commensal microorganisms within a host, comprising viruses, bacteria, fungi and eukaryotic parasites. As mNGS, particularly total RNA sequencing (i.e. metatranscriptomics), enables the identification of diverse and divergent viral sequences, it has been widely utilised for virus discovery [4-8].

Although the data generated by mNGS is bountiful and cost-effective, it comes with several inherent limitations, central of which is the possibility of reagent contamination [9]. Indeed, the contamination of mNGS data can be problematic when identifying microbes in the context of disease association and creates issues when attempting to identify the true host of a

novel microbe. The experimental preparation of samples for sequencing necessarily involves treatment with a variety of reagents, many of which have been shown to carry contaminating nucleic acids, including viral sequences [10-15]. Previous work has illuminated the extent of viral contamination in commonly used laboratory components, particularly those with small single-stranded (ss) DNA genomes [9, 14, 16-18]. Accordingly, there is a clear need for appropriate controls when characterizing novel viruses from metagenomic data. For example, metagenomic analysis of human plasma samples revealed the presence of sequences of Kadipiro virus, a double-stranded positive-sense RNA virus [19, 20]. However, the presence of these sequences was not confirmed via PCR, suggesting that they were contaminant in origin [19, 20]. An additional complication is that reagent-associated viral sequences are often not shared nor widespread across samples, only appearing intermittently [9].

Although mNGS has identified many novel viruses, diverse species of circular replication-associated protein encoding (CRESS) ssDNA viruses have been particularly prominent [21-25]. However, as noted above, ssDNA viruses, particularly CRESS viruses and their relatives including circoviruses, are common contaminants of reagents, leading to incorrect inferences of host associations [9, 26]. As well as DNA viruses, a variety of other microbial sequences are present in laboratory reagents, including bacteria, RNA viruses, and eukaryotic parasites [9, 27-30].

To further explore the diversity of contaminant sequences in laboratory components, particularly those derived from viruses, we used metatranscriptomics to investigate seven libraries of blank RNA sequencing samples representing sterile water extractions and library preparation reagents.

5. Methods

When generating total RNA sequencing libraries, we regularly utilise negative or ‘blank’ samples as experimental controls to assess the extent of reagent contamination. These controls are derived from extractions of the sterile water used at the elution step, and importantly, are expected to contain no nucleic acid material. In theory, these negative controls should generate no sequencing reads, however they can capture contamination during the DNA/RNA extraction or library preparation steps.

Herein, we analysed negative control sequencing libraries under different experimental conditions to identify likely contaminant sequences (**Table 1**). Total RNA was extracted using either the RNeasy Plus Universal Kit (Qiagen), RNeasy Plus Mini Kit (Qiagen) or the Total RNA purification Kit (Norgen BioTek Corp), as described in Table 1. RNA libraries were prepared with the Trio RNA-seq + UDI Library Preparation Kit (NuGEN) or the SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Clontech) and sequenced on the MiSeq, NextSeq or NovaSeq Illumina platforms, producing between 0.63Gb and 8.7Gb of data per library.

Analysis of virus-like sequences in laboratory reagents

Each sequencing library underwent trimming and *de novo* assembly of reads, completed using either the Trinity software with default settings [31] or MEGAHIT [32]. Sequence similarity searches using Diamond BLASTX were performed on the *de novo* assembled contigs against the GenBank non-redundant (nr) database [33, 34]. Specifically, we used a combination of e-value, hit length, and percentage similarity to determine the potential of a contig to be a viral sequence. The abundance of reagent-associated reads was calculated by comparing the number of contig reads to the total number of library reads (via mapping trimmed reads back to the contigs) as performed in previous studies [5, 8].

After initial identification, all potential contaminant sequences were subjected to phylogenetic analysis. To ensure high quality amino acid sequence alignments, only conserved sequence contigs that were >800 bp (>200 amino acids) in length were used in downstream analysis. Reference proteins including the highly conserved replicase, DNA polymerase and RNA-dependent RNA polymerase (RdRp) proteins were downloaded from the NCBI RefSeq database (**Table 2**). Contig and reference proteins were aligned using the L-INS-I algorithm in MAFFT v7 [35], with ambiguously aligned regions removed using Gblocks [36] which resulted in final sequence alignments of between 150-1000 amino acids in length (**Table 2**). Phylogenetic trees of all alignments were then estimated using the maximum likelihood method in IQ-TREE [37], using the model testing option and bootstrap resampling with 500 replications.

6. Results

In total, we identified 14 reagent-associated viral sequences in the negative (blank) control samples tested, including seven CRESS-like viral sequences, four novel *Tombusviridae*-like viral sequences, and single *Lentivirus*-like and *Totiviridae*-like viral sequences.

The abundance of reads in each library was calculated to compare the percentage of reads associated with viruses (**Figure 1**). This revealed that the virus-associated contigs identified were predominantly CRESS-like (**Figure 1b-e**). The L5 library only contained one virus-associated contig, associated with *Escherichia coli* phage PhiX 174 DNA: this was intentionally added into the sequencing run to add complexity and improve signal in the library. Both the L4 and L6 libraries did not contain long (>800bp) virus-associated contigs.

Novel reagent-associated virus-like sequences were identified in four of the seven libraries (**Table 3**). Seven novel circo-like viruses (termed Reagent-associated CRESS-like virus 1-7), four novel tombusvirus-like viruses (termed Reagent-associated tombus-like virus 1-4), and one totivirus-like and lentivirus-like sequence (termed Reagent-associated toti-like virus and Reagent-associated lenti-like virus, respectively) were identified in the L1, L2 and L3 libraries. The contigs ranged from 828-3878 bp in length and comprised 0.004-9.66% of reads in their associated libraries.

Because of the extensive genetic diversity within the *Circoviridae* we inferred two separate sequence alignments and hence two phylogenetic trees, representing the CRESS viruses and circoviruses taken independently, although both were based on the Rep protein sequence (**Figure 3**). All seven of the novel reagent-associated circovirus-like sequences exhibited greater sequence similarity to the CRESS viruses, and therefore were included in the CRESS virus phylogeny and termed reagent-associated CRESS-like viruses 1-7. These viruses occupied diverse locations across the phylogeny, although they were closely related to some previously identified reagent-associated viruses: Avon-Heathcote estuary associated circular viruses, *Circoviridae* sp. subtypes, Dromedary stool-associated circular virus subtypes, and Sandworm circovirus [5, 9] (**Figure 2**). It is notable that the CRESS viruses analysed derive from a variety of environments, and there is no clear pattern according to the host species of sample origin, which is anticipated in the case of contaminant sequences. The seven novel CRESS-like viruses identified also varied in abundance in the L1 and L3 libraries (0.01-

9.66%). In contrast, a phylogenetic analysis of the Rep protein of other members of *Circoviridae* (**Table 2),** containing what we hypothesise are *bona fide* viruses, reveals a pattern of host-based clustering (**Figure 3**). In particular, this phylogeny was characterised by two distinct clades of circoviruses: circoviruses, associated with vertebrate hosts, and cycloviruses associated with invertebrates.

Aside from ssDNA viruses, we identified an additional seven novel reagent-associated viral sequences in the blank control libraries. The first of these was a novel lentivirus-like sequence that we then used in an alignment of the retroviral Pol protein (**Table 2**). A phylogenetic tree was inferred from the alignment and the novel reagent-associated lenti-like virus was shown to cluster closely to Equine infectious anaemia viruses (EIAV), although occupying a relatively long branch within this clade (**Figure 4**).

Similarly, we identified four novel tombus-like sequences in the blank control samples: these were termed Reagent-associated tombus-like virus 1-4. A sequence alignment of the RNA-dependent RNA polymerase (RdRp) protein was used to infer a phylogenetic tree of these tombusvirus-like sequences that are commonly associated with plants (**Table 2**). Three of the novel tombus-like viruses cluster together in the same divergent clade that falls basal to majority of the tombus-like viruses (**Figure 5**). Only two tombus-like virus sequences fall in more divergent positions – Wenzhou tombus-like virus 11 and Sclerotinia sclerotiorum umbra-like virus 1. As these were both identified in metatranscriptomic studies [8, 39] it is possible that they reflect reagent contamination, although Sclerotinia sclerotiorum umbra-like virus 1 was found in two samples of *Sclerotinia sclerotiorum* (a fungus) compatible with its status as a true mycovirus [39, 40]. Additionally, Plasmopara viticola lesion associated tombus-like virus 2, which is also suggested to be a mycovirus, falls nearby (**Figure 5**). This virus sequence falls basal to a clade within the broader tombusvirus tree that includes a variety of plant viruses, including Groundnut rosette virus, Carrot mottle virus and Tobacco mottle virus. Reagent-associated tombus-like virus 3 was identified in blank library L3 at a relatively high abundance (1% of total reads), although it had a shorter (1574 bp) and likely incomplete genome compared to most tombusviruses (~4-5 kb).

Finally, the remaining novel sequence was related to the totiviruses, a family of double-strand RNA viruses commonly associated with fungi. The novel totivirus-like sequence was termed

Reagent-associated toti-like virus. It was used in an alignment of the RdRp protein (**Table 2**), from which a phylogenetic tree was estimated (**Figure 6**). This revealed that the sequence appears to be related to Scheffersomyces segobiensis virus (83% amino acid identity) associated with the fungus *Scheffersomyces segobiensis.*

7. Discussion

Viral sequences, particularly those with single-stranded DNA genomes, have previously been associated with common laboratory components [9], and these contaminant viral sequences have sometimes led to erroneous disease associations [14, 17, 18, 41]. Herein, using a series of blank controls comprising sterile water and commonly used laboratory reagents, we identified a diverse range of viral sequences.

Few laboratory reagents appear to be entirely free from contamination, particularly by ssDNA viruses, predominantly circoviruses [5, 9, 26]. Indeed, approximately half of the viral sequences identified here were CRESS-like members of the *Circoviridae*. Unfortunately, high levels of sequence diversity prevented us from obtaining a meaningful alignment of the Rep protein for the novel CRESS-like virus sequences obtained here and known *Circoviridae*. Accordingly, we divided the family into sub-groups, termed here as "host-associated circoviruses" (**Figure 3**) and "CRESS and CRESS-like viruses" and performed phylogenetic analyses on each (**Figure 2**). Notably, in the "host-associated circovirus" phylogeny viruses clustered based on broad host species of origin. In contrast, within the "CRESS and CRESS-like" phylogeny, clades could not be defined based on specific hosts or environments, and while many samples were originally derived from marine- or faeces-associated environments, these sequences did not cluster together. Interestingly, however, one of viruses identified in this study, reagent-associated CRESS-like virus 4, is most closely related to Avon-Heathcote Estuary associated circular virus 3, previously identified as a reagent-associated virus [42]. In addition, the seven novel CRESS-like sequences identified here were related to previously identified reagent-associated viruses, including those identified by Asplund et al. (highlighted in blue, **Figure 2**) [9], as well as Sandworm circovirus similarly proposed to be a reagent contaminant [43]. This strongly suggests that all these sequences are likely associated with laboratory reagents.

It is therefore clear that CRESS-like viruses are common experimental reagent contaminants, with widespread reagent-associated sequences dispersed throughout the CRESS phylogeny. This, along with the range of CRESS viruses of undetermined host origin, create major difficulties in determining the origin of novel CRESS viruses. Although there have been many new members of *Circoviridae* characterized in recent years, particularly novel cycloviruses [5, 44, 45], we suggest that current and future characterizations of novel circovirus- and CRESS-like genomes should be completed cautiously with additional confirmation steps.

We also identified several tombusvirus-like sequences in this study, as well as a totivirus- and lentivirus-like sequence. The *Tombusviridae* are a family of single-strand positive-sense RNA viruses are usually associated with mosaic diseases in plants. We identified four novel tombusvirus-like sequences associated with laboratory reagents, calling into question the provenance of other novel tombusviruses identified in some meta-transcriptomic studies [46]. The identification of reagent-associated tombusvirus-like sequences suggests that additional care should be taken when characterizing novel tombusvirus sequences, particularly when associating novel or previously identified tombusviruses with a host or disease. Similarly, although the natural hosts of the *Totiviridae* are fungi, other *Totiviridae* are associated with human-infecting protozoa, such as *Trichomonasvirus* associated with *Trichomonas vaginalis* [47] and *Giardiavirus* that likely infects *Giardia lamblia* protozoa [48, 49]. The novel reagent-associated totivirus identified in this study is distantly related to known totiviruses. We recommend that caution be taken when identifying novel totiviruses, especially if they are related to reagent-associated toti-like virus.

Lentiviruses are a genus within the *Retroviridae* and well documented in a wide range of vertebrate species. The novel sequence identified in this study – reagent-associated lenti-like virus – is closely related to several known sequences of equine infectious anemia virus (EIAV) that cause the chronic disease, equine infectious anemia (EIA) in horses. EIAV is transmissible through bodily secretions [50, 51], and has been suggested to be vector-borne through biting flies [52]. Although the novel reagent-associated lenti-like virus was genetically distinct from known EIAV sequences, care should obviously be taken to ensure that any EIAV-like virus is a true viral infection rather than a reagent contaminant.

In sum, this study further highlights the extent of viral sequences in commonly used laboratory reagents [9], and the power of mNGS to monitor contamination in microbiological laboratories [53]. Although the source of these contaminants is unknown and needs further scrutiny, we tentatively suggest that viral vectors (for example, in the *Lentiviridae*) represent a likely source. Factors to consider when assessing the presence of reagent contaminants include genome coverage, read depth and distribution of read alignments across genomes, and that potential contaminant sequences are often only present at low abundance and in multiple libraries. Importantly, reagent-associated viruses are often more prevalent in sequencing reads than assembled contigs, emphasising the importance of careful assessment when relying on read data alone for characterizing novel viruses and other microbial genomes [9, 26]. Finally, our work highlights the importance of employing additional steps such as PCR or cell culture to confirm the presence of the pathogen after initial metagenomic identification [9, 26]. Clearly, sequencing negative controls, such as that using sterile water and reagent mix as performed here, should become normal procedure in quality control.

8. Author statements

8.1 Authors and contributors

Conceptualization, E.C.H.; methodology, A.F.P, J.C, C.L and J.- S.E.; formal analysis, A.F.P.; writing—original draft preparation, A.F.P. and E.C.H..; writing—review and editing, A.F.P., E.C.H., C.L, J.C, J.-S.E.; funding acquisition, E.C.H.

8.2 Conflicts of interest

The authors declare that there are no conflicts of interest.

8.3 Funding information

This research was funded by an Australian Research Council Australian Laureate Fellowship to E.C.H (grant FL170100022).

8.4 Ethical approval

Not applicable.

8.5 Acknowledgements

We acknowledge the University of Sydney high performance computing cluster Artemis and Sydney Informatics Hub which was used for the analyses in this study.

9. References

1. **Grubaugh ND, Ladner JT, Lemey P, Pybus OG, Rambaut A, Holmes EC, et al.** Tracking virus outbreaks in the twenty-first century. *Nat Microbiol*. 2019;4:10-9.

2. **Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al.** Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*. 2020;395:565-74.

3. **Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al.** Spread of SARS-CoV-2 in the Icelandic Population. *New Eng J Med*. 2020; 382:2302-15.

4. **Zhang Y-Z, Chen Y-M, Wang W, Qin X-C, Holmes EC.** Expanding the RNA virosphere by unbiased metagenomics. *Annu Rev Virol*. 2019;6:119-39.

5. **Porter AF, Pettersson JHO, Chang W-S, Harvey E, Rose K, Shi M, et al.** Novel hepaci- and pegi-like viruses in native Australian wildlife and non-human primates. *Virus Evol*. 2020;6:veaa064.

6. **Geoghegan JL, Di Giallonardo F, Cousins K, Shi M, Williamson JE, Holmes EC.** Hidden diversity and evolution of viruses in market fish. *Virus Evol*. 2018;4:vey031.

7. **Harvey E, Rose K, Eden JS, Lo N, Abeyasuriya T, Shi M, et al.** Extensive diversity of RNA viruses in Australian ticks. *J Virol*. 2019;93; e01358-18.

8. **Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, et al.** Redefining the invertebrate RNA virosphere. *Nature*. 2016;540:539-43.

9. **Asplund M, Kjartansdóttir KR, Mollerup S, Vinner L, Fridholm H, Herrera JA, et al.** Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. *Clin Microbiol Infect*. 2019;25:1277-85.

10. **Kjartansdóttir KR, Friis-Nielsen J, Asplund M, Mollerup S, Mourier T, Jensen RH, et al.** Traces of ATCV-1 associated with laboratory component contamination. *Proc Natl Acad Sci USA*. 2015;112:E925-E6.

11. **Laurence M, Hatzis C, Brash DE.** Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS ONE*. 2014;9:e97876.

12. **Friis-Nielsen J, Kjartansdóttir KR, Mollerup S, Asplund M, Mourier T, Jensen RH, et al**. Identification of known and novel recurrent viral sequences in data from multiple patients and multiple cancers. *Viruses*. 2016;8:53.

13. **Lysholm F, Wetterbom A, Lindau C, Darban H, Bjerkner A, Fahlander K, et al.** Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. *PLoS ONE*. 2012;7:e30875.

14. **Smuts H, Kew M, Khan A, Korsman S.** Novel hybrid parvovirus-like virus, NIH-CQV/PHV, contaminants in silica column-based nucleic acid extraction kits. Journal of *Virology*. 2014;88:1398.

15. **Lusk RW.** Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing Data. *PLoS ONE*. 2014;9:e110808.

16. **Knox K, Carrigan D, Simmons G, Teque F, Zhou Y, Hackett J, et al.** No evidence of murine-like gammaretroviruses in CFS patients previously identified as XMRV-infected. *Science*. 2011;333:94-7.

17. **Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, et al.** The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J Virol*. 2013;87:11966-77.

18. **Paprotka T, Delviks-Frankenberry KA, Cingöz O, Martinez A, Kung H-J, Tepper CG, et al.** Recombinant origin of the retrovirus XMRV. *Science*. 2011;333:97-101.

19. **Ngoi CN, Siqueira J, Li L, Deng X, Mugo P, Graham SM, et al.** The plasma virome of febrile adult Kenyans shows frequent parvovirus B19 infections and a novel arbovirus (Kadipiro virus). *J Gen Virol*. 2016;97:3359-67.

20. **Ngoi CN, Siqueira J, Li L, Deng X, Mugo P, Graham SM, et al.** Corrigendum: the plasma virome of febrile adult Kenyans shows frequent parvovirus B19 infections and a novel arbovirus (Kadipiro virus). *J GenVirol*. 2017;98:517.

21. **Kerr M, Rosario K, Baker CCM, Breitbart M.** Discovery of four novel circular single-stranded DNA viruses in fungus-farming termites. *Microbiol Resour Ann*. 2018;6:e00318-18.

22. **Kazlauskas D, Dayaram A, Kraberger S, Goldstien S, Varsani A, Krupovic M.** Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. *Virology*. 2017;504:114-21.

23. **Krupovic M, Ghabrial SA, Jiang D, Varsani A**. *Genomoviridae*: a new family of widespread single-stranded DNA viruses. *Arch Virol*. 2016;161:2633-43.

24. **Rosario K, Breitbart M, Harrach B, Segales J, Delwart E, Biagini P, et al.** Revisiting the taxonomy of the family *Circoviridae*: establishment of the genus *Cyclovirus* and removal of the genus *Gyrovirus*. *Arch Virol*. 2017;162:1447-63.

25. **Varsani A, Krupovic M.** *Smacoviridae*: a new family of animal-associated single-stranded DNA viruses. *Arch Virol*. 2018;163:3213-4.

26. **Holmes EC.** Reagent contamination in viromics: all that glitters is not gold. *Clin Microbiol Infect*. 2019;25:1167-8.

27. **de Goffau MC, Lager S, Salter SJ, Wagner J, Kronbichler A, Charnock-Jones DS, et al**. Recognizing the reagent microbiome. *Nat Microbiol*. 2018;3:851-3.

28. **Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al.** Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol*. 2014;12:87.

29. **Zinter M, Mayday M, Ryckman K, Jelliffe-Pawlowski L, DeRisi J.** Towards precision quantification of contamination in metagenomic sequencing experiments. *Microbiome*. 2019;7:1-5.

30. **Stinson LF, Keelan JA, Payne MS.** Identification and removal of contaminating microbial DNA from PCR reagents: impact on low-biomass microbiome analyses. *Lett Appl Microbiol*. 2019;68:2-8.

31. **Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al.** Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech*. 2011;29:644-52.

32. **Li D, Liu C-M, Luo R, Sadakane K, Lam T-W.** MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 2015;3:1674-6.

33. **Buchfink B, Xie C, Huson DH.** Fast and sensitive protein alignment using DIAMOND. *Nat Meth*. 2015;12:59-60.

34. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.** BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.

35. **Katoh K, Standley DM.** MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability. *Mol Biol Evol*. 2013;30:772-80.

36. **Castresana J.** Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 2000;17:540-52.

37. **Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ.** IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015;32:268-74.

38. **Langmead B, Salzberg SL**. Fast gapped-read alignment with Bowtie 2. *Nat Meth*. 2012;9:357-9.

39. **Marzano S-YL, Nelson BD, Ajayi-Oyetunde O, Bradley CA, Hughes TJ, Hartman GL, et al**. Identification of diverse mycoviruses through metatranscriptomics characterization of the viromes of five major fungal plant pathogens. *J Virol*. 2016;90:6846-63.

40. **Mu F, Xie J, Cheng S, You MP, Barbetti MJ, Jia J, et al.** Virome characterization of a collection of *S. sclerotiorum* from Australia. *Front. Microbiol*. 2017;8:2540.

41. **Erlwein O, Robinson MJ, Dustan S, Weber J, Kaye S, McClure MO.** DNA extraction columns contaminated with murine sequences. *PLoS ONE*. 2011;6:e23484.

42. **Dayaram A, Goldstien S, Arguello-Astorga GR, Zawar-Reza P, Gomez C, Harding JS, et al.** Diverse small circular DNA viruses circulating amongst estuarine molluscs. *Infect Genet Evol*. 2015;31:284-95.

43. **Porter AF, Shi M, Eden J-S, Zhang Y-Z, Holmes EC.** Diversity and evolution of novel invertebrate DNA viruses revealed by meta-transcriptomics. *Viruses*. 2019;11:1092.

44. **Rosario K, Dayaram A, Marinov M, Ware J, Kraberger S, Stainton D, et al.** Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). *J Gen Virol*. 2012;93:2668-81.

45. **Islam SU, Lin W, Wu R, Lin C, Islam W, Arif M, et al.** Complete genome sequences of three novel cycloviruses identified in a dragonfly (Odonata: Anisoptera) from China. *Arch Virol*. 2018;163:2569-73.

46. **Culley AI, Lang AS, Suttle CA.** Metagenomic analysis of coastal RNA virus communities. *Science*. 2006;312:1795-8.

47. **Goodman RP, Ghabrial SA, Fichorova RN, Nibert ML.** *Trichomonasvirus*: a new genus of protozoan viruses in the family *Totiviridae*. *Arch Virol*. 2011;156:171-9.

48. **Wang AL, Yang HM, Shen KA, Wang CC.** Giardiavirus double-stranded RNA genome encodes a capsid polypeptide and a gag-pol-like fusion protein by a translation frameshift. *Proc Natl Acad Sci USA*. 1993;90:8595-9.

49. **Wang AL, Wang CC.** Viruses of the protozoa. *Annu Rev Microbiol*. 1991;45:251-63.

50. **Sellon DC, Fuller FJ, McGuire TC.** The immunopathogenesis of equine infectious anemia virus. *Virus Res*. 1994;32:111-38.

51. **Issel CJ, Adams WV, Jr., Meek L, Ochoa R**. Transmission of equine infectious anemia virus from horses without clinical signs of disease. *J Am Vet Med Assoc*. 1982;180:272-5.

52. **Hawkins JA, Adams WV, Jr., Wilson BH, Issel CJ, Roth EE.** Transmission of equine infectious anemia virus by Tabanus fuscicostatus. *J Am Vet Med Assoc*. 1976;168:63-4.

53. **Xiao Y, Zhang L, Yang B, Li M, Ren L, Wang J.** Application of next generation sequencing technology on contamination monitoring in microbiology laboratory. *Biosafety Health*. 2019;1:25-31.

10. Figures and tables

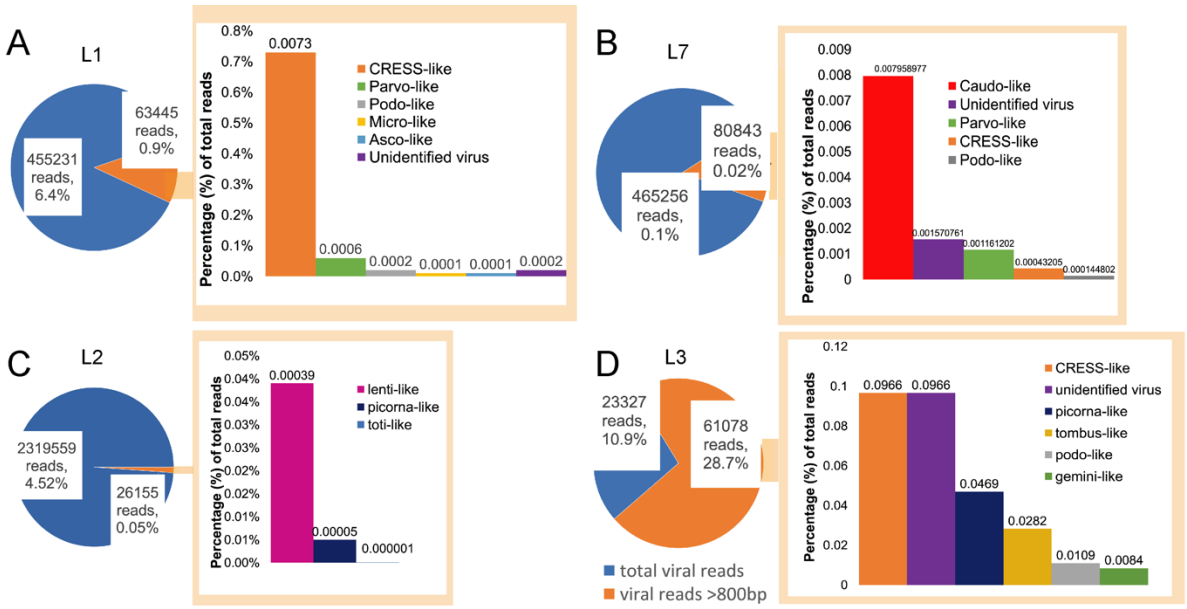

Figure 1. Abundance of viral reads in libraries L1, L2, L3, and L7. (A-D) Visual representation of the virus-associated reads in respective libraries, with the pie chart depicting the total number of long (>800 bp) virus-associated contigs (orange) compared to all the virus-associated reads (blue). The bar chart on the right denotes the proportion of contigs of associated with different virus families in the respective libraries.

Figure 2. Phylogenetic relationships of CRESS (ssDNA) viruses, including the seven novel CRESS-like viruses identified here and highlighted in red (reagent-associated CRESS-like viruses 1-7). Reagent-associated sequences determined previously are highlighted in blue. The clades that included the novel CRESS-like viruses identified here (A, B and G) are magnified on the right. The tree and other clades (C, D, E and F) are shown in higher resolution in Supplementary Figure 1. The tree was mid-point rooted for clarity purposes only. Bootstrap values greater than 70% are represented by asterisks next to nodes. All horizontal branch lengths are scaled according to number of amino acid substitutions per site.

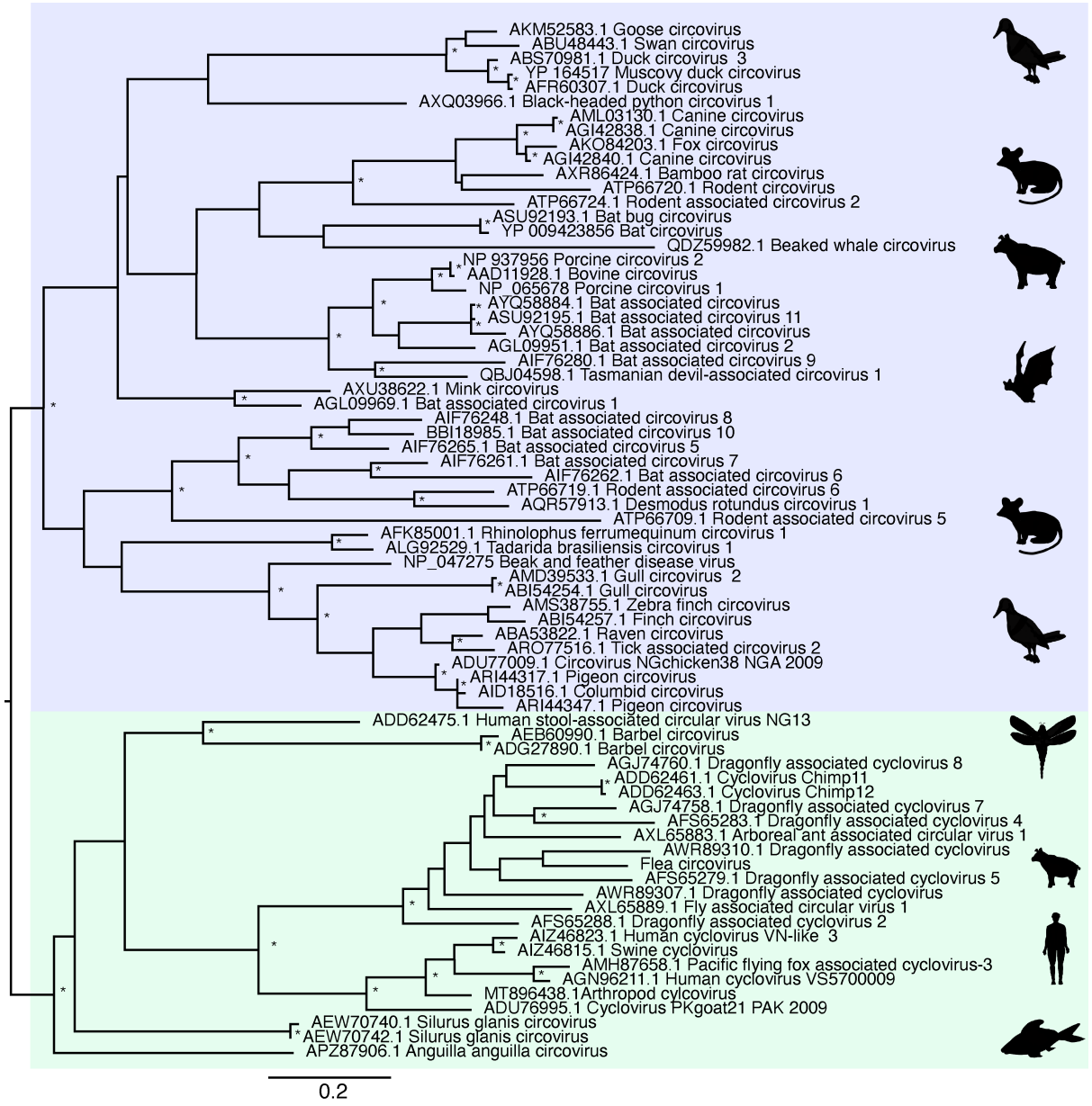

Figure 3. Phylogenetic relationships of ssDNA virus family C*ircoviridae,* based on hypothesised "host-associated" circoviruses. The tree has two major clades, comprising the circovirus clade (highlighted in blue), associated with vertebrate hosts, and the cyclovirus clade (highlighted in green), previously associated with invertebrate hosts. For clarity, the tree is mid-point rooted. Bootstrap values greater than 70% are represented by asterisks next to nodes. All horizontal branch lengths are scaled according to number of amino acid substitutions per site.

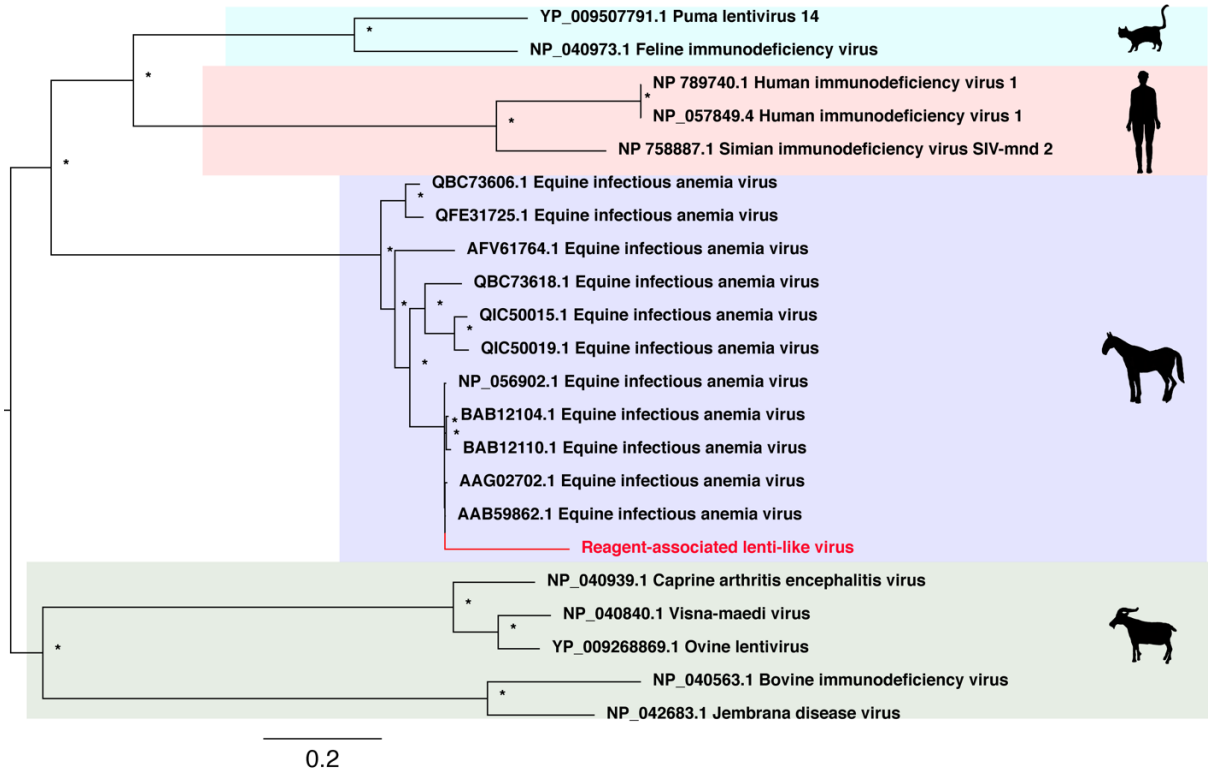

Figure 4. Phylogenetic relationships of RNA virus family *Lentiviridae* including the novel virus identified in this study, the novel sequence reagent-associated lenti-like virus. This virus is highlighted in red and falls within the Equine infectious anemia virus clade.

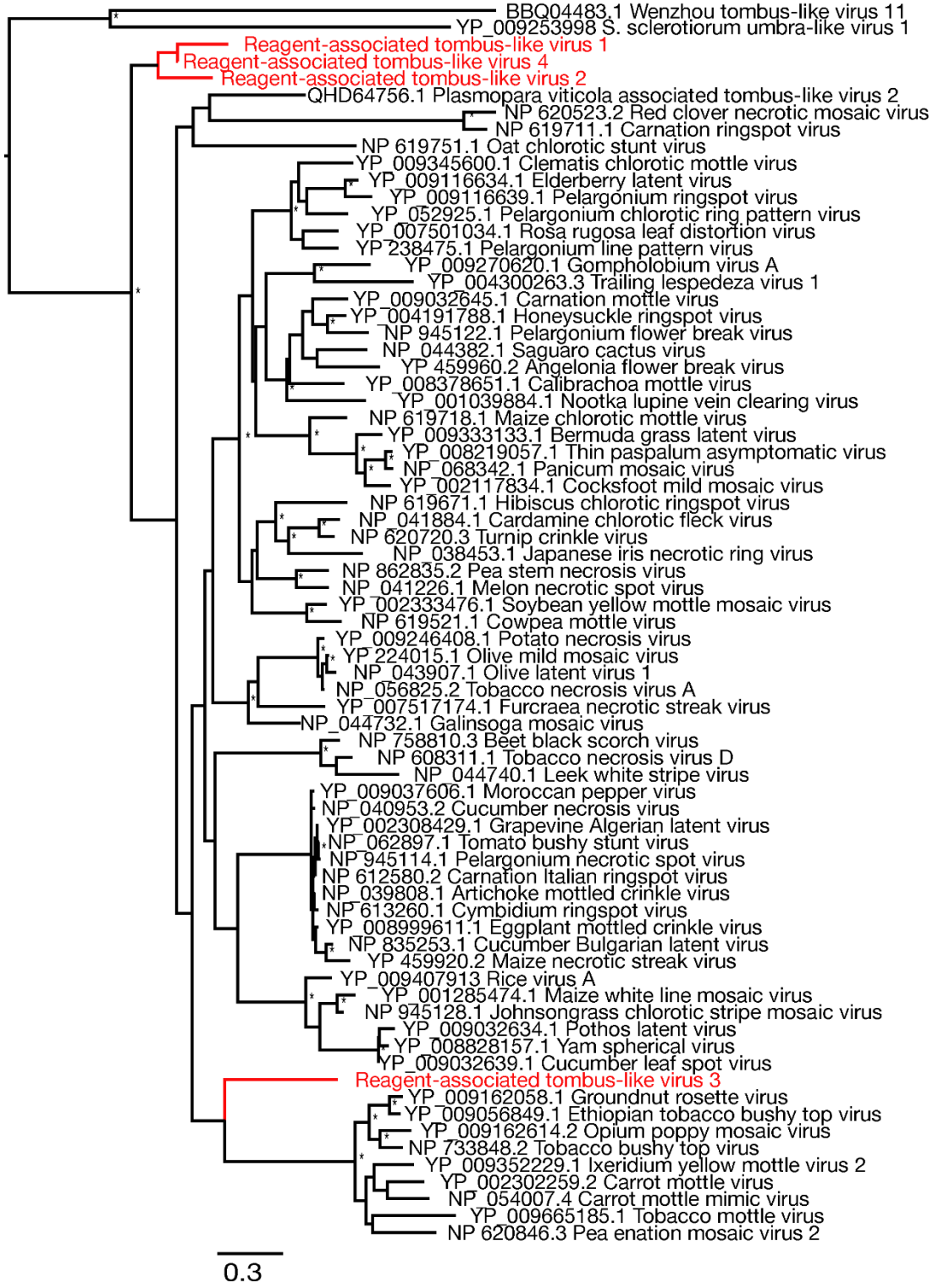

Figure 5. Phylogenetic relationships of RNA virus family *Tombusviridae* including the seven novel viruses identified in this study (highlighted in red). The phylogeny was mid-point rooted for clarity purposes only. Bootstrap values greater than 70% are represented by asterisks next to nodes. All horizontal branch lengths are scaled according to number of amino acid substitutions per site.

Figure 6. Phylogenetic relationships of RNA virus family *Totiviridae*, including the novel virus identified in this study - Reagent-associated toti-like virus (highlighted in red). For clarity, the tree was mid-point rooted. Bootstrap values greater than 70% are represented by asterisks next to nodes. All horizontal branch lengths are scaled according to number of amino acid substitutions per site.

471 **Table 1.** Experimental conditions of each blank negative control sample utilised here.

472

Library name	Sequencing platform	RNA extraction	Library preparation	Data generated	Library accession[1]
L1	Illumina Novaseq 6000 150 cycle kit (2x75nt reads)	RNeasy Plus Universal Kits (Qiagen)	Trio RNA-seq +UDI (NuGEN)	11,940,824 paired reads (1.8Gb)	SRR14737471
L2	Illumina Novaseq 6000 150 cycle kit (2x75nt reads)	RNeasy Plus Universal Kits (Qiagen)	Trio RNA-seq +UDI (NuGEN)	57,606,392 paired reads (8.7Gb)	SRR14737470
L3	Illumina MiSeq, v3 150 cycle kit (2x75nt reads)	RNeasy Plus Mini Kit (Qiagen)	SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Clontech)	4,156,504 paired reads (0.63 Gb)	SRX6803604
L4	Illumina NextSeq 500, mid-output 150 cycle kit (2x75nt reads)	Total RNA Purification Kit (Norgen Biotek)	SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Clontech)	32,279,914 paired reads (4.91 Gb)	SRR14737469
L5	Illumina MiSeq 150 cycle kit (2x75nt reads)	Total RNA purification Kit (Norgen BioTek Corp)	SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Clontech)	7,342,876 paired reads (1.10 Gb)	SAMN20355437
L6	Illumina MiSeq 150 cycle kit (2x75nt reads)	Total RNA purification Kit (Norgen BioTek Corp)	SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Clontech)	10,978,253 paired reads (1.65 Gb)	SAMN20355438
L7	Illumina MiSeq 150 cycle kit (2x75nt reads)	Total RNA purification Kit (Norgen BioTek Corp)	SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Clontech)	8,564,269 1.28 Gb	SRR14737466

[1]The sequencing data for each library can be accessed via the sequence read archive (SRA) using the associated accession numbers.

473

474 **Table 2.** Reference proteins for each sequence alignment performed in this analysis.

Reference protein	Reference acronym	Taxonomy	Number of sequences in analysis	Alignment length (amino acid, AA)
Viral replicase protein	Rep	CRESS	221	672 AA
Viral replicase protein	Rep	*Circoviridae*	69	161 AA
Polymerase peptide	Pol	*Lentiviridae*	11	478 AA
RNA-dependent RNA polymerase	RdRp	*Totiviridae*	95	125 AA
RNA-dependent RNA polymerase	RdRp	*Tombusviridae*	87	256 AA

475

476 **Table 3.** Novel reagent-associated viral sequences identified in this study.

Virus name	Accession	Abundance in library (%) of total reads, rRNA removed)	Length (bp)	Library
Reagent-associated tombus-like virus 1	MZ824229	1.28	1204	L3
Reagent-associated tombus-like virus 2	MZ824228	0.46	828	L3
Reagent-associated tombus-like virus 3	MZ824227	1.08	1574	L3
Reagent-associated tombus-like virus 4	MZ824226	1.29	1410	L3
Reagent-associated toti-like virus	MZ824225	0.001	920	L2
Reagent-associated lenti-like virus	MZ824230	0.004	962	L2
Reagent-associated CRESS-like virus 1	MZ824237	0.78	3878	L1
Reagent-associated CRESS-like virus 2	MZ824236	0.24	2377	L1
Reagent-associated CRESS-like virus 3	MZ824235	0.02	1592	L1
Reagent-associated CRESS-like virus 4	MZ824234	2.89	2663	L3
Reagent-associated CRESS-like virus 5	MZ824233	9.66	3027	L3
Reagent-associated CRESS-like virus 6	MZ824232	4.98	3517	L3
Reagent-associated CRESS-like virus 7	MZ824231	0.01	1124	L1

477