Discovery of three RNA viruses using ant transcriptomic datasets

Eleni Kleanthous1 · Ingrida Olendraite1 · Nina I. Lukhovitskaya1 · Andrew E. Firth1

Received: 10 August 2018 / Accepted: 20 October 2018 / Published online: 10 November 2018 © The Author(s) 2018

Abstract
Three novel RNA viruses, named Formica fusca virus 1 (GenBank accession no. MH477287), Lasius neglectus virus 2 (MH477288) and Myrmica scabrinodis virus 2 (MH477289), were discovered in ants collected in Cambridge, UK. The proposed virus names were given based on the hosts in which they were identified. The genome sequences were obtained using de novo transcriptome assembly of high-throughput RNA sequencing reads and confirmed by Sanger sequencing. Phylogenetic analysis showed that Formica fusca virus 1 grouped within the family Nyamiviridae, Lasius neglectus virus 2 grouped within the family Rhabdoviridae and Myrmica scabrinodis virus 2 belongs to the family Dicistroviridae. All three viruses are highly divergent from previously sequenced viruses.

Insect-infecting viruses comprise a large and diverse group, but their full diversity has as yet been poorly investigated [1]. Further investigation may provide new insights into the long-term evolution and origin of viruses and may also reveal new tools for molecular biology research and biotechnology. Also, the host range of some of these viruses may make them suitable to use as biological control agents against invasive or pestiferous insect species.

We generated RNA-Seq libraries from ants sampled from various locations in Cambridge, UK [2]. Sequencing data have been deposited in ArrayExpress (http://www.ebi.ac.uk/arrayexpress) under the accession number E-MTAB-5781. Trimmed RNA-Seq reads were assembled de novo into contigs using TRINITY v2.3.2 [3, 4], and assembled contigs were compared using BLASTX [5] against a database of all RNA virus protein sequences obtained from GenBank as described previously [2]. Although we focused on genetically characterising three of the viral sequences with highest coverage, our analysis of Trinity assemblies revealed that further uncharacterised virus sequences were present in the RNA-Seq datasets. Primers were designed based on the three viral sequences, and overlapping fragments were amplified by PCR and sequenced by the Sanger method. Apart from the first 202 nt of FfusV-1 and the terminal 59 nt of MsaV-2, sequences of the entire genomes were obtained by Sanger sequencing; 100% nucleotide identity was confirmed between assembled and amplified sequences.

To determine the taxonomy of the three viruses, related sequences were identified within GenBank (5 Dec 2017, taxID = 10239 “viruses”), using the amino acid sequence of the entire ORF encoding the RNA-dependent RNA polymerase (RdRp) as the query in TBLASTN. Reference (RefSeq database) and selected non-reference (nr/nt database) sequences with > 25% coverage match were used for phylogenetic analysis. Amino acid sequences of the RdRp-encoding ORF were aligned using MUSCLE (v3.8.31) [6], and phylogenetic trees were inferred using MRBAYES v3.2.6 [7], with one million generations and default parameters. Phylogenetic trees were visualized with FIGTREE v1.4.3.

To determine genome organisations, potential viral ORFs were identified using GETORF from the EMBOSS package [8], and each translated ORF was queried with HHpred [9] against the Pfam [10] and PDB [11] databases and identified based on homology to previously annotated proteins.

The sequence of Formica fusca virus 1 (FfusV-1) comprises 9,851 nt. The genome organization of the virus follows a typical Mononegavirales gene order with five ORFs, encoding proteins in the sequence 3′ – N (nucleocapsid, 361 aa) – P (polymerase cofactor, 184 aa) – M (matrix protein, 149 aa) – G (glycoprotein, 527 aa) – L (RdRp, 1877 aa) – 5′
in negative polarity (Fig. 1). The 5′ leader and 3′ trailer of the antigenome are of lengths 104 and 120 nt, respectively. Identification of the proteins encoded by ORFs 2 and 3 was uncertain as HHpred e-values were > 1, but they were predicted to be P and M based on their genome position. Genome coverage was obtained to a mean depth of 28 for the negative strand and 7 for the positive strand, consistent with active virus replication. Phylogenetic analysis indicated that FfusV-1 is most closely related to Orinoco virus and suggested it belongs in the genus *Orinovirus* within the family *Nyamiviridae* (Fig. 2).

The sequence of Lasius neglectus virus 2 (LneV-2) comprises 12,041 nt. The 5′ leader and 3′ trailer sequences of the antigenome are of 115 and 174 nt, respectively, and share nine nucleotides of terminal complementarity. LneV-2 follows a typical *Mononegavirales* genome organisation, encoding proteins in the same order as FfusV-1, N (470 aa), P (433 aa), M (248 aa), G (506 aa) and L (2124 aa) (Fig. 1). As with FfusV-1, the proteins encoded by ORFs 2 and 3 were not confidently predicted by HHpred but inferred based on position in the genome. Genome coverage was obtained to a mean depth of 45 for the negative strand and 14 for the positive strand, again consistent with active virus replication. Phylogenetic analysis indicated that LneV-2 belongs in the family *Rhabdoviridae* but does not belong to any of the defined genera within the family; instead it falls within an unclassified cluster of arthropod-infecting rhabdoviruses. Based on the phylogenetic tree, LneV-2 is most closely related to Hubei rhabdo-like virus 1, *Spodoptera frugiperda* rhabdovirus and Wuhan ant virus (Fig. 2).

The sequence of Myrmica scabrinodis virus 2 (MsaV-2) comprises 10,663 nt. It contains two ORFs of 5553 nt and 4089 nt in length separated by an intergenic sequence of 196 nt and flanked by 5′ and 3′ UTRs of 630 and 199 nt plus the poly(A) tail, respectively (Fig. 1). HHpred analysis indicated that ORF 1 encodes the non-structural proteins helicase, protease and RdRp, whereas ORF 2 encodes three jelly-roll structural capsid proteins. Genome coverage was obtained to a mean depth of 262 for the positive strand and 0.2 for the negative strand. Importantly, the negative-strand reads – although sparse – were distributed throughout the genome, indicating the probable presence of a full-length negative strand. Phylogenetic analysis indicated that this virus belongs in the genus *Triatovirus* within the family *Dicistroviridae* of the order *Picornavirales* (Fig. 2). In dicistroviruses, the second ORF starts at a non-AUG initiation site due to the presence of an unusual intergenic region internal ribosomal entry site (IGR-IRE), which directs translation of the 3′ ORF via an initiator-Met-tRNA independent mechanism [12]; a typical IGR-IRE structure was predicted in MsaV-2 and used to infer the non-AUG start site of ORF 2.
Discovery of RNA viruses using ant transcriptomic datasets

**A**  

**Bornaviridae**  

- KM114265.1: Lovelidges garter snake virus 1  
- JX065207.1: Parrot bornavirus 1  
- LN13680.1: Vanegated squirrel bornavirus 1  
- A3J31524.1: Borna disease virus 2  
- U04608.1: Borna disease virus 1  
- LC120625.1: Parrot bornavirus 5  
- KJ756399.2: Aquatic bird bornavirus 2  
- KC464478.1: Canary bornavirus 2  
- KC464471.1: Canary bornavirus 1

**Nyamaviridae**  

- KX884457.1: Wenling crustacean 12  
- KM817644.1: Wenzhou crab virus 1  
- KX884436.1: Wenzhou tapeworm virus 1  
- HM849038.1: Soybean cyst nematode midway virus  
- KX884406.1: Belhai rhabdovirus virus 6  
- KX884405.1: Belhai rhabdovirus 6  
- KX884404.1: Belhai rhabdovirus 5

**Crustaviridae**  

- KX884401.1: Belhai rhabdovirus 4

**Nyamaviridae**  

- KX884407.1: Belhai rhabdovirus 5  
- KX884408.1: Belhai rhabdovirus 3

**Xinmoviridae**

- KX884428.1: Hebei rhabdovirus virus 7  
- KX148553.1: Gamble virus  
- MF176247.1: Culex mononega-like virus 2  
- KM817638.1: Shuangao fly virus 2  
- KR822819.1: Drosophila unisigna virus 1

**B**  

**Lysavirus**  

- M13215.1: Rabies virus  
- EU293109.1: Lages bat virus  
- KJ689309.3: Drosophila melanogaster sigma virus  
- GG410979.1: Drosophila obscura sigma virus  
- GG29441.1: Triborggar virus  
- KM885010.1: Yata virus  
- GU212866.1: Chandipura virus  
- EU735657.1: Cocal virus Indiana 2  
- KJ703428.1: Vesicular stomatitis Indiana virus  
- AY840978.1: Tapela virus  
- A174498.1: Snakehead rhabdovirus  
- LH0883.1: Infectious hematopoietic necrosis virus  
- A065187.1: Rice yellow stunt virus  
- GU734660.1: Potato yellow dwarf virus  
- KX636164.1: Physostigia chlorotic mottle virus  
- GN806056.1: Eggplant mottled dwarf virus  
- L32603.1: Sonchus yellow net virus  
- KY704866.1: Citrus chlorotic spot virus  
- EF687878.1: Lettuce yellow mottle virus  
- A1909584.2: Lettuce necrotic yellows virus  
- KM817851.1: Wuhan insect virus 5  
- AB735628.2: Persimmon virus  
- KM884447.1: Hebei rhabdovirus 1

**Sigaviridae**  

- KM817645.1: Wahan ant virus  
- KM817631.1: Shuangao bedbug virus 2  
- KM817634.1: Saxia water strider virus 5  
- KR822813.1: Drosophila busskii rhabdovirus  
- KM817659.1: Wuhan mosquito virus 9  
- KX884456.1: Wenling crustacean virus 11  
- KJ735609.1: Drosophila longicauda rhabdovirus  
- KM817836.1: Shanyang fly virus 3  
- KM817641.1: Wuhan fly virus 5  
- KM817641.1: Wuhan house fly virus 2  
- KM822822.1: Drosophila sturtvanti rhabdovirus  
- KU754213.1: Wittyham virus  
- KR822811.1: Drosophila subobscura rhabdovirus  
- KC602379.1: Farmington virus
Acknowledgements This work was supported by a Wellcome Trust grant [106207] and a European Research Council grant [646891] to A.E.F.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval The research reported here did not involve any studies on human subjects or vertebrate animals performed by any of the authors.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Fig. 2 (continued)

Figures

References

1. Shi M, Lin XD, Tian JH et al (2016) Redefining the invertebrate RNA virosphere. Nature 540:539–543
2. Orendairaite I, Lukhovitskaya NI, Porter SD, Valles SM, Firth AE (2017) Polycipivirdae: a proposed new family of polycistronic picorna-like RNA viruses. J Gen Virol 98:2368–2378
3. Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol 29:644–652
4. Haas BJ, Papanicolaou A, Yassour M et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512
5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
6. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797
7. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574
8. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277
9. Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33:W244–W248

10. Finn RD, Coggill P, Eberhardt RY et al (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44:D279–D285

11. Berman HM, Westbrook J, Feng Z, Gililand G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242

12. Nakashima N, Uchiumi T (2009) Functional analysis of structural motifs in dicistroviruses. Virus Res 139:137–147