Sequence, assembly and count datasets of viruses associated to the pine processionary moth Thaumetopoea pityocampa (Denis & Schiffermüller) (Lepidoptera, Notodontidae) identified from transcriptomic high-throughput sequencing

Franck Dorkeld, Réjane Streiff, Laure Sauné, Guillaume Castel, Marie Helene Ogliastro, Carole Kerdelhué

Data in Brief 48 (2023) 109180

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Sequence, assembly and count datasets of viruses associated to the pine processionary moth *Thaumetopoea pityocampa* (Denis & Schiffermüller) (Lepidoptera, Notodontidae) identified from transcriptomic high-throughput sequencing

Franck Dorkeld[a], Réjane Streiff[a], Laure Sauné[a], Guillaume Castel[a], Mylène Ogliastro[b], Carole Kerdelhué[a,*]

[a] *CBGP, INRAE CIRAD IRD Institut Agro Université de Montpellier, 755 avenue du Campus Agropolis, CS30016, Montferrier-sur-Lez cedex F-34988, France*
[b] *DGIMI, INRAE Université de Montpellier, Place Eugène Bataillon, 34095 Montpellier cedex 5, France*

A R T I C L E   I N F O

Article history:
Received 2 March 2023
Revised 19 April 2023
Accepted 20 April 2023
Available online 25 April 2023

Dataset link: Transcriptomic sequence data for the pine processionary moth Thaumetopoea pityocampa and alignments used for the phylogenetic analyses of the cypovirus, the iflavirus and the densovirus identified (Original data)

Keywords:
Transcriptome
High-throughput sequencing
De novo assembly
In silico virus identification
Pine processionary moth
Phylogeny

A B S T R A C T

The pine processionary moth *Thaumetopoea pityocampa* is a Lepidopteran pest species occurring in the Western Mediterranean. It causes heavy pine defoliations and it is a public and animal health concern because of its urticating caterpillars. Very little is known about the viruses associated to this species, as only two viruses were described so far. We here present a dataset corresponding to 34 viral transcripts, among which 27 could be confidently assigned to 9 RNA and DNA viral families (*Iflaviridae, Reoviridae, Partitiviridae, Permutotetraviridae, Flaviviridae, Rhabdoviridae, Parvoviridae, Baculoviridae* and *PolyDNAviridae*). These transcripts were identified from an original transcriptome assembled for the insect host, using both blast search and phylogenetic approaches. The data were acquired from 2 populations in Portugal and 2 populations in Italy. The transcripts were de novo assembled and used to identify viral sequences by homology searches. We also provide information about the populations and life

[*] Corresponding author.
E-mail address: Carole.Kerdelhue@inrae.fr (C. Kerdelhué).

https://doi.org/10.1016/j.dib.2023.109180

stages in which each virus was identified. The data produced will allow to enrich the virus taxonomy in Lepidopteran hosts, and to develop PCR-based diagnostic tools to screen colonies across the range and determine the distribution and prevalence of the identified viral species.

Specifications Table

Subject	Virology
Specific subject area	Virus identification from original transcriptomic data
Type of data	Fastq (raw read sequences) Assemblies : de novo assembled transcriptome for the host *Thaumetopoea pityocampa* and 34 viral transcripts Fasta: alignments used for phylogenetic analyses Table: raw reads counts per viral transcript in each population and life stage Table: alignment statistics and hit annotations of transcripts showing significant BLAST hits with reference viral sequences Figures: phylogenetic trees and position of the Cypovirus, the Iflavirus and the Densovirus identified in this study
How the data were acquired	RNA-seq data were sequenced on an Illumina HiSeq2000 platform using the 2 × 100 bp paired-end protocol. Twenty-seven libraries were built, corresponding to 4 populations (2 in Italy and 2 in Portugal) and 6 to 8 life stages for each population.
Data format	Raw: raw read sequences Analyzed: list of annotated viral transcripts identified, alignments, read counts
Description of data collection	Transcriptomic data were assembled using Trinity2.0.2 and merged using CD-HIT-EST4.5.5 and CAP3 02/10/15. Assembled transcripts were first subjected to a Blastx against the reference viral RefSeq database of the NCBI, and the transcripts that were successfully aligned to a viral reference were subjected to a second Blastx search against the NCBI nr database. Virus identification was based on homology results. To identify the relative contributions of the different RNA libraries (corresponding to various populations and/or life stages) to the transcripts finally identified as viral sequences, we used Bowtie2 with default parameters to align the cleaned reads against the final set of candidate viral transcripts identified in the assembled PPM transcriptome and used the number and percentage of reads originating from each library as metrics. To further characterize the best assembled virus genomes, we conducted phylogenetic analyses for the Cypovirus, the Densovirus and the Iflavirus.
Data source location	• Mata Nacional de Leiria • Country: Portugal • Latitude and longitude: 39°47′ N, 8°58′ W • Cimolais • Country: Italy • Latitude and longitude: 12°27′ E, 46°19′ N • Tregnano • Country: Italy • Latitude and longitude: 11°09′ E, 45°30′ N
Data accessibility	Repository name for raw sequences: NCBI BioProject Data identification number: PRJNA663237 Direct URL to data: http://www.ncbi.nlm.nih.gov/bioproject/663237 Repository name for 34 viral transcripts: NCBI Genbank Data identification number: accession numbers MT796426-MT796428, MT799182-MT799183, MW584279-MW584285, MW584288-MW584291, MW584293, MW584296, MW584298- MW584302, MW584206-MW584316

(*continued on next page*)

	Repository name for the assembled transcriptome, the alignments used for phylogenetic analyses and the raw reads counts per viral transcript in each population and life stage: Institutional INRAE repository A Table containing the links to the raw data separately for each of the 27 librairies is also provided in the same repository. Direct link to data: 10.57745/AYQBQB

Value of the Data

- The data correspond to annotated viral transcripts associated to the Lepidopteran pest *Thaumetopoea pityocampa*, the pine processionary moth (PPM).
- The data significantly enrich the list of Lepidoptera-associated viruses and provide genomic data for future exploration of their diversity.
- These new sequences will serve as references to target these viruses, and will allow screening natural colonies of the PPM to determine their prevalence in the field.
- Targeting larvae showing signs of viral disease will allow identifying which virus are entomopathogenic and could be candidates for future management solutions.

1. Objective

The main objective of this work was to explore the virome associated to natural populations of *T. pityocampa* by mining large RNAseq datasets corresponding to all developmental stages, various phenologies and geographic origins. We first assembled a PPM transcriptome and developed bio-informatics analyses to specifically identify viral genomes and transcripts. This approach allowed to identify new viruses associated with the PPM and assign their origin. It was very important to expand the knowledge about the viruses associated to the PPM because only 2 exogenous and one endogenous viruses were described so far, from the sister species *T. wilkinsoni* [1]. We provide the list of identified viruses, their annotation and best blast-hit as well as the corresponding transcript sequences.

2. Data Description

We first assembled a PPM reference transcriptome from the raw RNAseq data generated from 27 libraries; this transcriptome was comprised of 198,336 transcripts corresponding to 88,121 unigenes. Concerning quality assessment, the Cegma procedure allowed recovering the 248 genes at full length (100%). The Busco approach also confirmed that completeness was very high, as it retrieved 100% of the 303 eukaryote genes at full length, 99.5% of the arthropod genes (1061 at full length, 3 fragmented and 2 missing genes) and 98.8% of the insect genes (1638 complete, 14 fragmented and 6 missing genes). The raw data corresponding to this transcriptome were deposited in the NCBI BioProject PRJNA663237 and the transcriptome assembly is available in the INRAE institutional dataverse at 10.57745/AYQBQB. Table 1 shows the number of raw and cleaned paired reads per population and developmental stage (27 libraries), and proportion of mapped reads on the assembled transcripts.

The bioinformatic procedure described below then allowed to identify and annotate 34 viral transcripts. Among those, 27 were assigned to 9 virus families and 7 remained unclassified due to poor taxonomic information available in the public databases. Table 2 summarizes the characteristics and annotation of each viral transcript and gives the corresponding NCBI accession number and Blast results. RNA viruses were the most represented, with 6 virus families, namely *Iflaviridae, Reoviridae, Partitiviridae, Permutotetraviridae, Flaviviridae* and *Rhabdoviridae*. We also found DNA viruses from 3 families, both with small (*Parvoviridae*) and large genomes (*Baculoviridae* and *PolyDNAviridae*). The numbers of viral reads obtained from the different transcriptome

Table 1
Number of raw and cleaned paired reads per population and developmental stage, and proportion of mapped reads on the assembled transcripts.

Population Library code	Stage	# raw read pairs	# cleaned read pairs	% reads mapped back to transcripts
Cimolais (Italy)				
1	L1 larvae	38,661,411	35,564,006	81.32%
2	L2 larvae	41,021,841	37,709,113	82.52%
3	L3 larvae	41,010,324	37,713,232	83.61%
4	L4 larvae	13,690,903	12,643,194	84.42%
5	L5 larvae	50,736,491	44,626,888	84.59%
6	Early pupae	42,449,326	39,187,603	84.62%
7	Late pupae	54,739,835	50,323,960	84.52%
8	Adults	34,251,971	31,602,186	81.41%
Tregnano (Italy)				
9	L1 larvae	51,314,703	47,158,804	82.96%
10	L2 larvae	29,668,151	27,188,282	84.86%
11	L3 larvae	24,789,340	22,615,028	83.41%
12	L4 larvae	33,608,031	30,912,034	84.23%
13	L5 larvae	29,293,135	26,804,959	84.07%
14	Early pupae	40,739,299	37,669,684	83.95%
15	Adults	55,893,599	51,181,710	81.14%
Leiria SP (Portugal)				
16	Eggs	38,803,348	34,478,275	81.42%
17	L3 larvae	40,550,188	36,300,947	83.74%
18	L5 larvae	43,752,664	40,660,297	91.70%
19	Early pupae	55,936,647	52,278,035	90.75%
20	Late pupae	57,463,974	53,488,169	86.74%
21	Adults	43,248,551	40,265,025	87.61%
Leiria WP (Portugal)				
22	Eggs	41,464,384	37,438,856	79.08%
23	L3 larvae	35,699,902	31,709,969	85.06%
24	L5 larvae	44,360,149	41,117,253	91.38%
25	Early pupae	43,469,212	40,440,752	89.86%
26	Late pupae	56,184,645	52,026,117	89.99 %
27	Adults	55,480,253	51,785,133	87.08%

libraries can be found in the INRAE repository (10.57745/AYQBQB), and allow to determine to which sampling site and life stage each virus was associated.

The three figures we provide show the phylogenetic placements of the Cypovirus, the Iflavirus and the Densovirus we identified in the present work.

Fig. 1 showed that the Cypovirus identified in the present study is distantly related to the TpCPV5 described in Ref. [1] from the sister host species *T. wilkinsoni*. It branches as a sister group close to Cypovirus 2 and Cypovirus 16, in the same clade as Cypovirus 4.

Fig. 2 shows the phylogenetic tree of Iflavirus. The virus we identified from *T. pityocampa* belongs to the main Iflavirus lineage including the type species "Infectious flacherie virus" and did not fall in the same clade as the TpIV1 identified in Ref. [1] but appeared closely related to Iflavirus species associated to parasitoids (namely *Venturia canescens* picorna-like virus, *Dinocampus coccinellae* paralysis virus and *Lysiphlebus fabarum* Iflavirus).

The phylogeny shown in Fig. 3 shows the monophyly of the genus Iteradensovirus, which formed a strongly supported clade differentiated from the Ambidensoviruses included in the analysis. As we could include the 5 Iteradensovirus species recognized by the ICTV, we conclude that the Iteradensovirus we identified from *T. pityocampa* grouped with Lepidoptera Iteradensovirus 2 (from *Casphalia castanea*) together with 5 viruses isolated either from Lepidopterans (*Danaus plexippus* or *Sibine fusca*) or from bat and bird feces, or from a plant (*Hordeum marinum*). This group of 7 closely related sequences falls within a clade including also Lepidoptera iteradensovirus 1 and 4. The identity levels were above 85% when comparing the 7 NS1 cited

Table 2
Transcripts showing similarity to viral references. Library # refers to the codes given in Table 1. VP: viral protein; NSP: non-structural protein; CPV: Cypovirus; RdRp: RNA-dependent RNA polymerase. The genome range size for viruses in each family is given between brackets.

Virus	Trancript ID	Seq. length	GenBank accession number	Blast hit description	Alignment length (% identity)	e-value	obs. in library #
Iflavirus (8.8-9.7 kb)	TR140100|c0_g1_i1	608 bp	MW584290	polyprotein, *Lysiphlebus fabarum* RNA virus, *Iflaviridae*	203 aa (47%)	7.51e-64	19
	TR143551|c0_g1_i1	448 bp	MW584291	polyprotein, partial, Bee iflavirus 1	141 aa (44%)	6.04e-28	19
	TR73151|c0_g2_i1	1098 bp	MW584279	polyprotein, *Lysiphlebus fabarum* RNA virus type A, *Iflaviridae*	369 aa (52%)	1.25e-118	19
	TR73151|c1_g1_i1	309 bp	MW584293	polyprotein, *Lysiphlebus fabarum* RNA virus, *Iflaviridae*	107 aa (44%)	2.18e-22	19
	TR10271|c1_g1_i1	518 bp	MW584296	RdRp, *Venturia canescens* picorna-like virus	172 aa (67%)	1.77e-79	19
Cypovirus (CPV) (25 kb, segmented genome; 10 segments 1-4 kb)	TR63295|c0_g2_i1	3690 bp	MW584298	VP3, *Erinnyis ello* CPV2	1168 aa (34%)	0.0	24,25,26,27
	TR90152|c0_g1_i1	3360 bp	MW584280	VP4, *Erinnyis ello* CPV2	1134 aa (34%)	9.64e-174	24,25,26,27
	TR92789|c0_g2_i1	3697 bp	MW584281	RdRp, *Erinnyis ello* CPV2	1238 aa (44%)	0.0	24,25,26,27
	TR93454|c0_g1_i1	2112 bp	MW584299	VP5, *Inachis io* CPV2	622 aa (31%)	2.52e-61	24,25,26,27
	TR98576|c0_g1_i1	2042 bp	MW584300	Unknown protein, *Choristoneura occidentalis* CPV16	288 aa (33%)	5.15e-36	24,25,26,27
	TR77729|c0_g1_i1	3597 bp	MW584301	Unknown protein, *Choristoneura occidentalis* CPV16	1206 aa (34%)	5.17e-177	24,25,26,27

(*continued on next page*)

Table 2 (*continued*)

Virus	Trancript ID	Seq. length	GenBank accession number	Blast hit description	Alignment length (% identity)	e-value	obs. in library #
Rhabdovirus (11-15 kb)	TR101106|c2_g3_i1	1514 bp	MW584302	polymerase, Yinshui bat virus	265 aa (61%)	6.70e-125	All
	TR103332|c1_g2_i2	2222 bp	MW584282	RdRp, partial, *Muscina stabulans* sigmavirus	141 aa (60%)	1.87e-46	All
	TR104746|c0_g1_i1	1028 bp	MW584283	L protein, *Orgyia pseudotsugata* Orgi virus	145 aa (38%)	1.49e-21	All
	TR105009|c0_g2_i1	5765 bp	MW584284	RdRp, *Bactrocera dorsalis* sigmavirus	322 aa (58%)	0.0	All
Betapartitivirus (6 kb, segmented genome; 3 segments 1.7-2.2 kb)	TR105233|c0_g1_i1	2164 bp	MT799183	Coat protein, *Plasmopara viticola* lesion associated Partitivirus 7	593 aa (30%)	2.16e-70	22
	TR85286|c0_g1_i1	926 bp	MW584306	Coat protein, *Plasmopara viticola* lesion associated Partitivirus 7	197 aa (37%)	2.93e-30	22
	TR85286|c1_g1_i1	1214 bp	MW584285	Capsid protein, *Heterobasidion* partitivirus 8	410 aa (33%)	5.19e-42	22
	TR91956|c0_g1_i1	2210 bp	MT799182	RdRp, Soybean thrips partiti-like virus 8	696 aa (56%)	0.0	22
Iteradensovirus (5 kb)	TR17180|c0_g1_i1	2310 bp	MT796426	Non structural protein NS1, Iteravirus sp from bat guano	750 aa (98%)	0.0	5,16,17,22,23
	TR46092|c0_g1_i1	2040 bp	MT796427	Structural protein VP, Iteravirus sp from bat guano	679 aa (93%)	0.0	5,16,17,22,23
Betabaculovirus (80-180 kb)	TR5412|c0_g1_i1	439 bp	MW584307	Hypothetical protein AsGV069, *Agrotis segetum* granulovirus	143 aa (42%)	5.02e-27	5,11,13,16
	TR81448|c0_g1_i1	1178 bp	MW584308	Hypothetical protein AsGV069, *Agrotis segetum* granulovirus	208 aa (43%)	1.17e-42	1,2,4,5,7,11,13,16,17

(*continued on next page*)

Table 2 (*continued*)

Virus	Trancript ID	Seq. length	GenBank accession number	Blast hit description	Alignment length (% identity)	e-value	obs. in library #
PolyDNAvirus (6-20 kb)	TR95017|c0_g1_i1	2490 bp	MW584309	N-gene1, *Hyposoter didymator* ichnovirus	415 aa (51%)	7.62e-107	All
Permutotetravirus (5.6 kb)	TR46077|c0_g1_i1	5621 bp	MT796428	Capsid protein precursor, *Thosea asigna* alphapermutotetravirus	555 aa (49%)	1.79e-161	23
Flavivirus (10-11 kb)	TR50290|c0_g1_i1	368 bp	MW584310	Putative polyprotein, *Lampyris noctiluca* flavivirus 1	123 aa (49%)	2.07e-25	22
	TR50290|c1_g1_i1	610 bp	MW584311	Putative polyprotein, *Lampyris noctiluca* flavivirus 1	202 aa (55%)	1.23e-61	22
Unclassified viruses	TR98942|c7_g1_i1	2054 bp	MW584288	Polymerase Acidic protein, Lepidopteran orthomyxo-related virus OKIAV178	543 aa (43%)	3.32e-99	All but 4
	TR62311|c0_g1_i1	1261 bp	MW584312	Glycoprotein, Hymenopteran phasma-related virus OKIAV229	412 aa (33%)	5.26e-69	26
	TR12337|c0_g2_i1	407 bp	MW584313	Putative RdRp, Notori virus	134 aa (46%)	6.16e-29	22
	TR26798|c0_g1_i1	356 bp	MW584314	Glycoprotein precursor, Kaisodi virus	119 aa (37%)	1.06e-19	22
	TR93735|c0_g1_i1	2723 bp	MW584315	Putative RdRp, Raphidiopteran tombus-related virus	430 aa (50%)	7.92e-134	1,2,4-6,9,11,13,16,17,23
	TR93735|c2_g1_i1	785 bp	MW584316	RdRp, *Diaphorina citri* associated C-virus	213 aa (53%)	2.47e-68	11,13,16
	TR80667|c2_g1_i1	975 bp	MW584289	Glycoprotein, Hymenopteran phasma-related virus OKIAV244	313 aa (52%)	4.70e-90	22

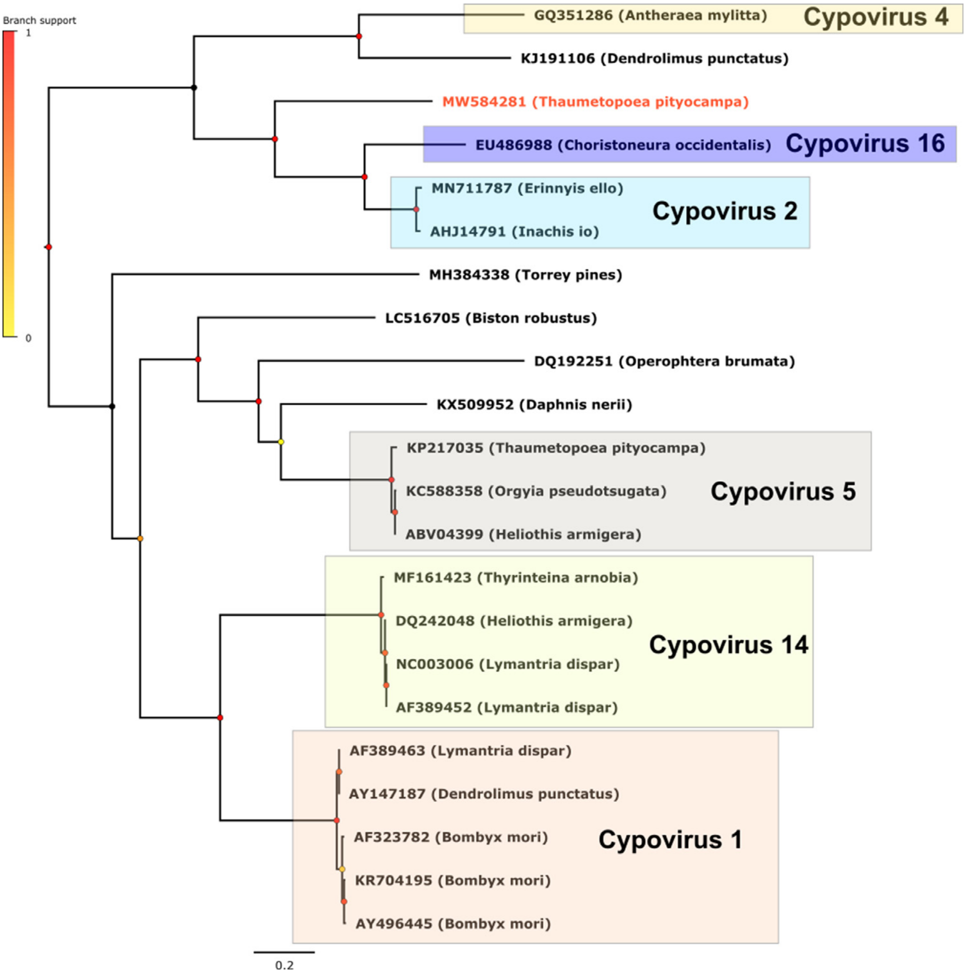
Branch support
1
0
GQ351286 (Antheraea mylitta)
Cypovirus 4
KJ191106 (Dendrolimus punctatus)
MW584281 (Thaumetopoea pityocampa)
EU486988 (Choristoneura occidentalis)
Cypovirus 16
MN711787 (Erinnyis ello)
AHJ14791 (Inachis io)
Cypovirus 2
MH384338 (Torrey pines)
LC516705 (Biston robustus)
DQ192251 (Operophtera brumata)
KX509952 (Daphnis nerii)
KP217035 (Thaumetopoea pityocampa)
KC588358 (Orgyia pseudotsugata)
ABV04399 (Heliothis armigera)
Cypovirus 5
MF161423 (Thyrinteina arnobia)
DQ242048 (Heliothis armigera)
NC003006 (Lymantria dispar)
AF389452 (Lymantria dispar)
Cypovirus 14
AF389463 (Lymantria dispar)
AY147187 (Dendrolimus punctatus)
AF323782 (Bombyx mori)
KR704195 (Bombyx mori)
AY496445 (Bombyx mori)
Cypovirus 1
0.2

Fig. 1. Cypovirus phylogeny reconstructed by Maximum Likelihood method (midpoint rooted tree). For each virus, the name of the host is indicated in parenthesis. The main clades of Cypovirus are indicated and the newly identified cypovirus of this study is written in red. Scale bar indicates number of amino acid changes per site and aLRT branch support values at each node are indicated by a color code.

above, and even reached 98% between the transcript we identified in the present study and the virus identified in a bat guano in Croatia.

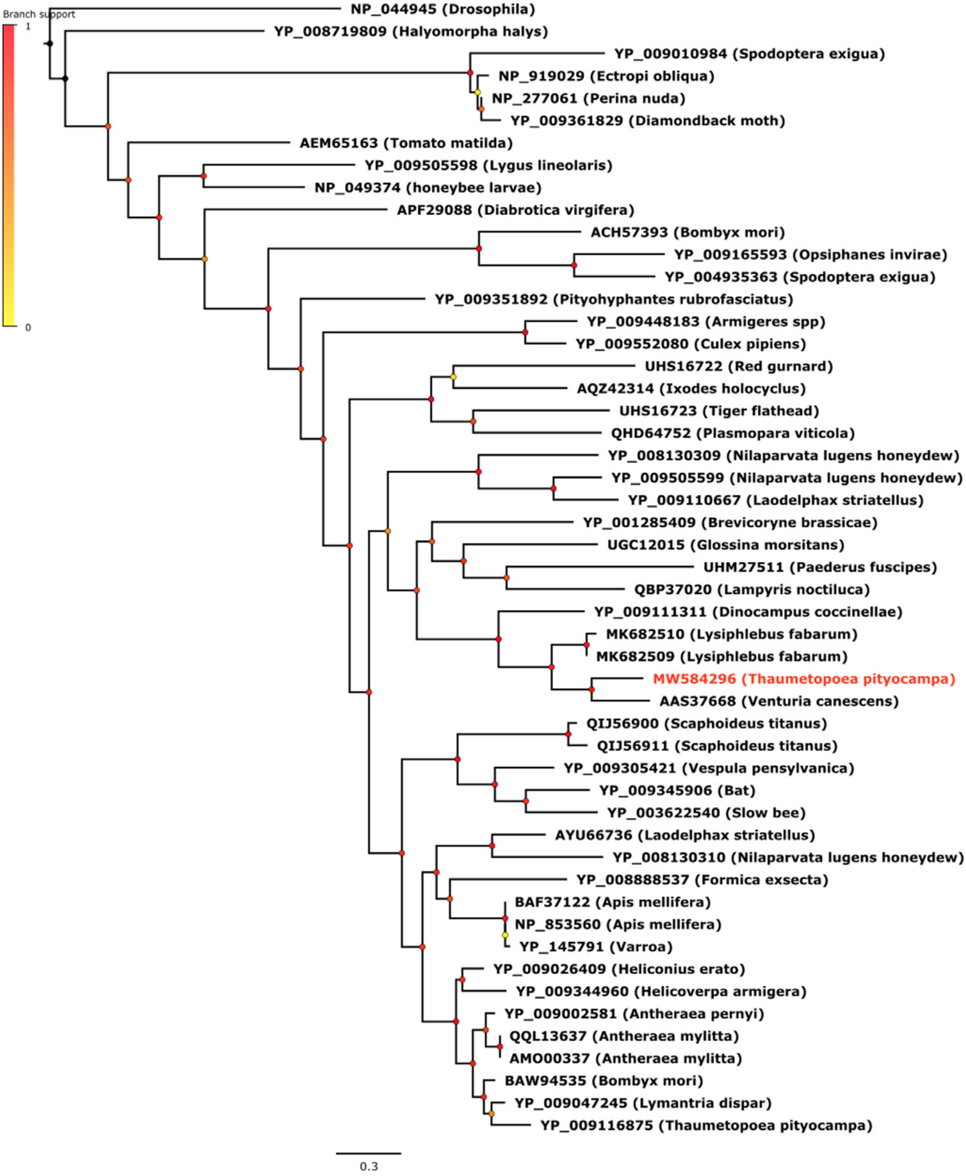
Branch support
1
0
NP_044945 (Drosophila)
YP_008719809 (Halyomorpha halys)
YP_009010984 (Spodoptera exigua)
NP_919029 (Ectropi obliqua)
NP_277061 (Perina nuda)
YP_009361829 (Diamondback moth)
AEM65163 (Tomato matilda)
YP_009505598 (Lygus lineolaris)
NP_049374 (honeybee larvae)
APF29088 (Diabrotica virgifera)
ACH57393 (Bombyx mori)
YP_009165593 (Opsiphanes invirae)
YP_004935363 (Spodoptera exigua)
YP_009351892 (Pityohyphantes rubrofasciatus)
YP_009448183 (Armigeres spp)
YP_009552080 (Culex pipiens)
UHS16722 (Red gurnard)
AQZ42314 (Ixodes holocyclus)
UHS16723 (Tiger flathead)
QHD64752 (Plasmopara viticola)
YP_008130309 (Nilaparvata lugens honeydew)
YP_009505599 (Nilaparvata lugens honeydew)
YP_009110667 (Laodelphax striatellus)
YP_001285409 (Brevicoryne brassicae)
UGC12015 (Glossina morsitans)
UHM27511 (Paederus fuscipes)
QBP37020 (Lampyris noctiluca)
YP_009111311 (Dinocampus coccinellae)
MK682510 (Lysiphlebus fabarum)
MK682509 (Lysiphlebus fabarum)
MW584296 (Thaumetopoea pityocampa)
AAS37668 (Venturia canescens)
QIJ56900 (Scaphoideus titanus)
QIJ56911 (Scaphoideus titanus)
YP_009305421 (Vespula pensylvanica)
YP_009345906 (Bat)
YP_003622540 (Slow bee)
AYU66736 (Laodelphax striatellus)
YP_008130310 (Nilaparvata lugens honeydew)
YP_008888537 (Formica exsecta)
BAF37122 (Apis mellifera)
NP_853560 (Apis mellifera)
YP_145791 (Varroa)
YP_009026409 (Heliconius erato)
YP_009344960 (Helicoverpa armigera)
YP_009002581 (Antheraea pernyi)
QQL13637 (Antheraea mylitta)
AMO00337 (Antheraea mylitta)
BAW94535 (Bombyx mori)
YP_009047245 (Lymantria dispar)
YP_009116875 (Thaumetopoea pityocampa)
0.3

Fig. 2. Iflavirus phylogeny reconstructed by Maximum Likelihood method (midpoint rooted tree). For each virus, the name of the host is indicated in parenthesis. The newly identified iflavirus of this study is written in red. Scale bar indicates number of amino acid changes per site and aLRT branch support values at each node are indicated by a color code.

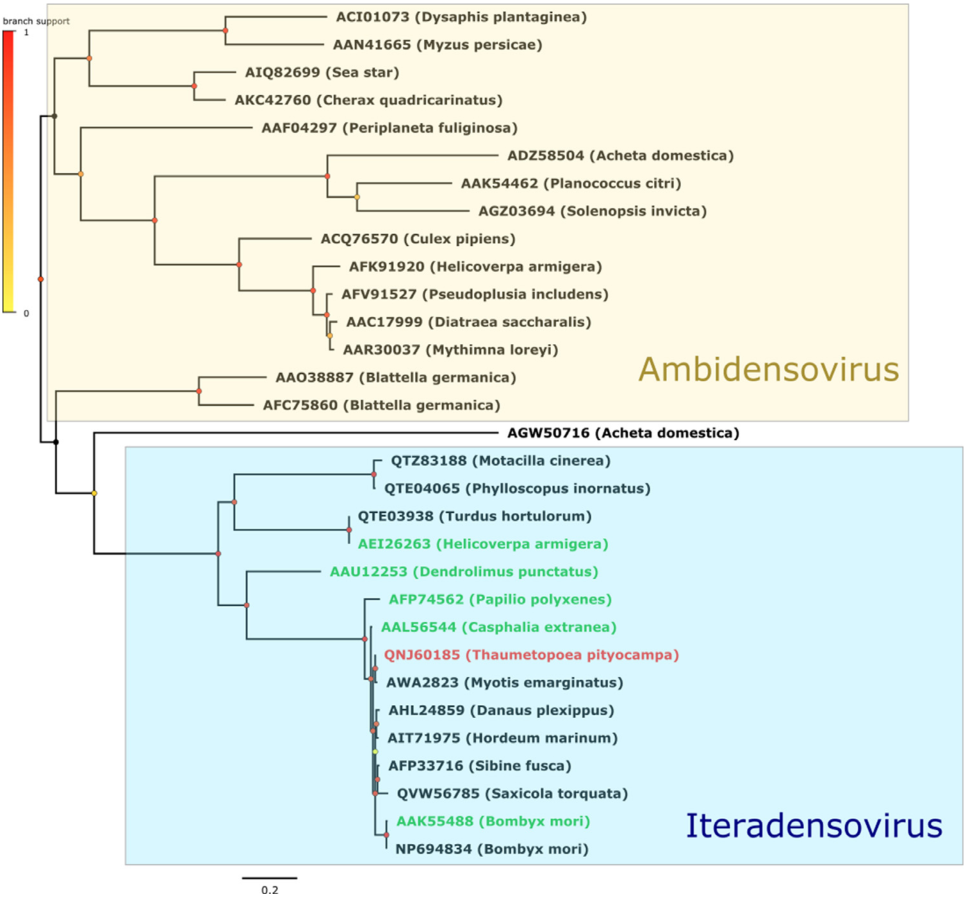
branch support
1
0
ACI01073 (Dysaphis plantaginea)
AAN41665 (Myzus persicae)
AIQ82699 (Sea star)
AKC42760 (Cherax quadricarinatus)
AAF04297 (Periplaneta fuliginosa)
ADZ58504 (Acheta domestica)
AAK54462 (Planococcus citri)
AGZ03694 (Solenopsis invicta)
ACQ76570 (Culex pipiens)
AFK91920 (Helicoverpa armigera)
AFV91527 (Pseudoplusia includens)
AAC17999 (Diatraea saccharalis)
AAR30037 (Mythimna loreyi)
AAO38887 (Blattella germanica)
AFC75860 (Blattella germanica)
Ambidensovirus
AGW50716 (Acheta domestica)
QTZ83188 (Motacilla cinerea)
QTE04065 (Phylloscopus inornatus)
QTE03938 (Turdus hortulorum)
AEI26263 (Helicoverpa armigera)
AAU12253 (Dendrolimus punctatus)
AFP74562 (Papilio polyxenes)
AAL56544 (Casphalia extranea)
QNJ60185 (Thaumetopoea pityocampa)
AWA2823 (Myotis emarginatus)
AHL24859 (Danaus plexippus)
AIT71975 (Hordeum marinum)
AFP33716 (Sibine fusca)
QVW56785 (Saxicola torquata)
AAK55488 (Bombyx mori)
NP694834 (Bombyx mori)
Iteradensovirus
0.2

Fig. 3. Densovirus phylogeny reconstructed by Maximum Likelihood method (midpoint rooted tree). For each virus, the name of the host is indicated in parenthesis. Ambidensovirus and Iteradensovirus are indicated. The newly identified iteradensovirus of this study is written in red while iteradensovirus reference strains available from ICTV are written in green. Scale bar indicates number of amino acid changes per site and aLRT branch support values at each node are indicated by a color code.

3. Experimental Design, Materials and Methods

3.1. Sampling and Raw Data Acquisition

Raw paired-end reads were obtained from Illumina HiSeq2000 RNA sequencing of individuals sampled from Italy (populations Cimolais, 12°27′ E, 46°19′ N; Tregnago, 11°09′ E, 45°30′ N) and Portugal (summer (SP) and winter (WP) populations from Leiria, 39°47′ N, 8°58′ W). The different developmental stages analyzed in each population are shown in Table 1. All molecular procedures were the same as described in Ref. [2].

3.2. Transcriptome Assembly and Quality Assessment

The raw reads were trimmed using Trimmomatic v.0.33 [3] using the following parameters: ILLUMINACLIP: adaptors_file.fa: 2: 40: 15; HEADCROP: 12; SLIDINGWINDOW: 4: 15 and MINLEN:

30 combined with Prinseq-lite v.0.20.2 [4] to eliminate polyA tails (parameters -trim_tail_left 5, -trim_ tail_right 5, -min_len 30, -out_format 3). The remaining reads were processed with Flash v.1.2.11 (Fast Length Adjustment of SHort reads) [5] to merge r1 and r2 reads.

The cleaned reads, obtained after the quality filtering step described above, were used for *de novo* transcript assembly using Trinity v.2.0.2 [6] with the normalization option on and default kmer value. To reduce redundancy, the transcripts were further merged into clusters using Cd-Hit-Est v.4.5.4 [7,8] with an identity threshold of 98 (parameters: -c 21 0.98 -l 20 -M 0 -B 1). These clusters were then analyzed with Cap3 v.02/10/15 [9] using parameters -o 200 -p 99. Finally, we excluded transcripts with low coverage by using the FPKM metric (Fragments Per Kilobase transcript length per million fragments Mapped) from the Rsem v.1.3.0 program [10]. Transcripts were removed when FPKM was lower than 1.

We mapped the filtered reads back to the assembled transcripts using Bowtie2 v.2.2.4 with default parameters [11] and we used the number of such reads as a quality assessment estimator. The completeness of the transcriptome assembly was assessed using the Cegma (Core Eukaryotic Genes Mapping Approach) pipeline v.2.5 [12], which searches for 248 orthologous groups of proteins [13]. We also used the Busco v.3.0 program [14,15] to search for 303 conserved eukaryotic genes, 1,066 conserved arthropod genes and 1,658 insect genes.

3.3. Bioinformatic Procedures for Virus Identification

The pipeline is summarized in Fig. 4. First, we searched for homology between *de novo* transcripts and the viral proteins present in the viral RefSeq database of NCBI (available at https://www.ncbi.nlm.nih.gov/refseq, version refseq_viral_proteic_2021-01-07) using the Blastx program [16] with an e-value threshold of 1.e-10 (parameters –outfmt 5; -gapopen 11; -gapextend 1; -word_size 3; -matrix BLOSUM62). Second, the subset of transcripts that could successfully be aligned against a viral protein in the viral RefSeq database and that were longer than 30 amino acids were subjected to a second Blastx search against the NCBI nr database (release 2021-06-16). They were retained only if this Blast search retrieved a viral protein as best Blast Hit in nr and if the alignment between the query and the subject was longer that 100 amino acids. We then restricted the results to the best alignments based on lowest E-values, highest bitscores and highest identities. Virus identification was based on homology results when Blast search provided consistent results. In some cases, the viral sequences found in the GenBank database originated from large metabarcoding programs that did not provide any taxonomic information. We then selected the most informative taxonomy within the search results when bit score and e-values were equivalent and the query start/end were equal.

To identify the relative contributions of the different RNA libraries (corresponding to various populations and life stages) to the viral transcripts, we used Bowtie2 with default parameters to align the cleaned reads against the final set of candidate viral transcripts. We used the number of reads originating from each library as metrics.

3.4. Phylogenetic Analyses

Analyses corresponding to the genus *cypovirus* were carried out on a set of aminoacid sequences of the RNA-dependent RNA polymerase (RdRp). The alignment included representative members of the genus recognized by the ICTV (6 of the 16 species), the putative Cypovirus 19 isolated from *Operophtera brumata*, 14 sequences identified in GenBank as corresponding to undescribed Cypovirus and the Cypovirus identified in the present study (TR92789, Genbank accession number MW584281). Analyses corresponding to the genus *iflavirus* were carried out on a dataset corresponding to the aminoacid sequences of the polyprotein of 50 representative members of the genus and the iflavirus identified in the present study (TR10271, Genbank accession number MW584296). Analyses corresponding to the genus *densovirus* were carried out on an

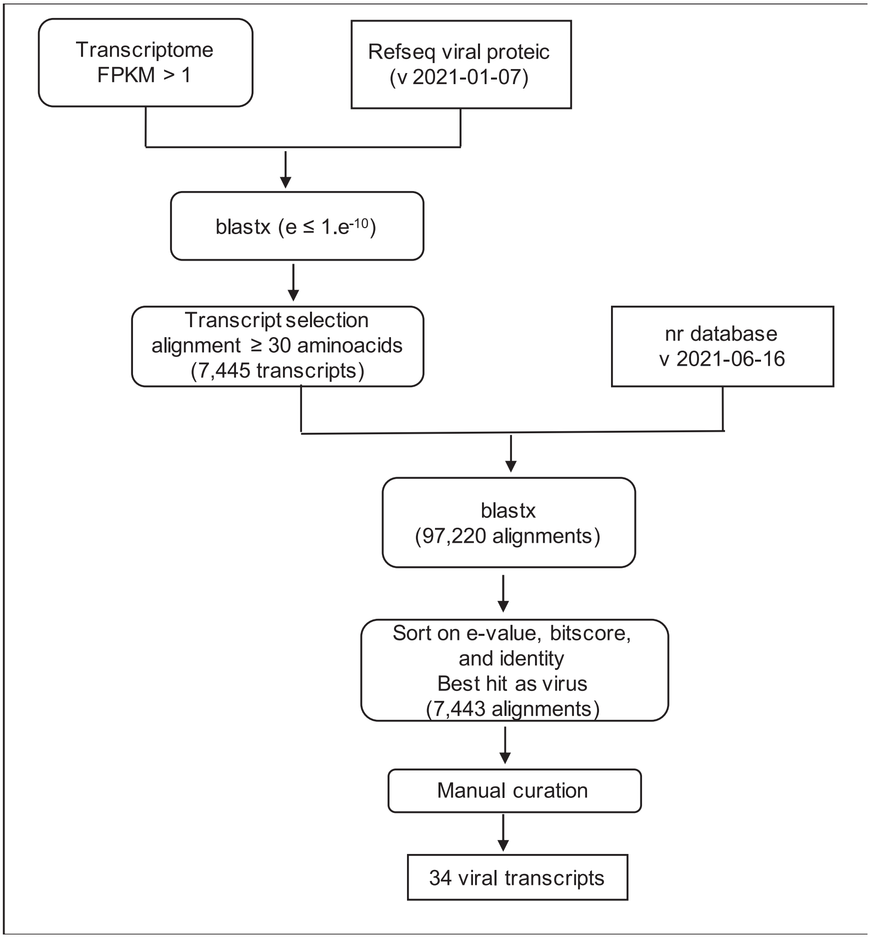

Fig. 4. Bio-informatics workflow developed to identify potential viral transcripts.

alignment of the aminoacid sequences of the NS1 protein of 30 representative members of the genus including the 5 Iteradensovirus species recognized by the ICTV and the iteradensovirus identified in the present study (TR17180, Genbank accession number QNJ60185).

For each dataset, multiple protein alignments were generated with the MAFFT v.7 alignment program [17] with default parameters, using a G-INS-i iterative refinement method. Gblocks method [18] implemented in SEAVIEW v5.0.4 [19] was used to eliminate poorly aligned positions and divergent regions, resulting in 690, 136 and 250 amino acids for respectively cypovirus, iflavirus and densovirus final sequence alignments. Optimal substitution models were identified using the SMS program [20] as the LG +G+I+F model (cypovirus and densovirus) or as the LG +G+I model (iflavirus).

Phylogenetic reconstruction was performed by a maximum likelihood (ML) approach in PHYML v3.0 [21] implemented in SEAVIEW v5.0.4, with a statistical approximate likelihood ratio test (aLRT) for branch support. Unrooted phylogenetic trees were visualized and edited with FIGTREE v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) using the midpoint rooting option.

Ethics Statements

This study did not involve any experiment conducted on human or vertebrate animals.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Transcriptomic sequence data for the pine processionary moth Thaumetopoea pityocampa and alignments used for the phylogenetic analyses of the cypovirus, the iflavirus and the densovirus identified (Original data) (Dataverse).

CRediT Author Statement

Franck Dorkeld: Software, Formal analysis, Data curation, Visualization; **Réjane Streiff:** Conceptualization, Writing – review & editing; **Laure Sauné:** Investigation, Resources; **Guillaume Castel:** Formal analysis, Writing – review & editing, Visualization; **Mylène Ogliastro:** Conceptualization, Writing – original draft, Supervision; **Carole Kerdelhué:** Conceptualization, Writing – original draft, Supervision, Project administration, Funding acquisition.

Acknowledgments

The authors are grateful to Manuela Branco and Susana Rocha (University of Lisbon, Portugal) and to Andrea Battisti and Laura Berardi (University of Padua, Italy) for providing the PPM samples. The authors would also like to thank the bioinformatics platforms ABiMS (Roscoff, France) for providing access to their computational resources.

This research was funded by the French National Agency for Research (ANR, project GenoPheno, grant ANR-10-JCJC-1705-01) and by a support by INRA, EFPA (ECODIV) department (projet innovant 2019 EVIL).

References

[1] A.K. Jakubowska, R. Nalcacioglu, A. Millán-Leiva, A. Sanz-Carbonell, H. Muratoglu, S. Herrero, Z. Demirbag, In search of pathogens: transcriptome-based identification of viral sequences from the pine processionary moth (*Thaumetopoea pityocampa*), Viruses 7 (2) (2015) 456–479, doi:10.3390/v7020456.

[2] B. Gschloessl, F. Dorkeld, H. Berges, G. Beydon, O. Bouchez, M. Branco, A. Bretaudeau, C. Burban, E. Dubois, P. Gauthier, E. Lhuillier, J. Nichols, S. Nidelet, S. Rocha, L. Sauné, R. Streiff, M. Gautier, C. Kerdelhué, Draft genome and reference transcriptomic resources for the urticating pine defoliator *Thaumetopoea pityocampa* (Lepidoptera: Notodontidae), Mol. Ecol. Res. 18 (3) (2018) 602–619, doi:10.1111/1755-0998.12756.

[3] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120, doi:10.1093/bioinformatics/btu170.

[4] R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets, Bioinformatics 27 (2011) 863–864, doi:10.1093/bioinformatics/btr026.

[5] T. Magoč, S. Salzberg, FLASH: Fast length adjustment of short reads to improve genome assemblies, Bioinformatics 27 (21) (2011) 2957–2963, doi:10.1093/bioinformatics/btr507.

[6] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol. 29 (7) (2011) 644–652, doi:10.1038/nbt.1883.

[7] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (23) (2012) 3150–3152, doi:10.1093/bioinformatics/bts565.

[8] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (13) (2006) 1658–1659, doi:10.1093/bioinformatics/btl158.
[9] X. Huang, A. Madan, CAP3: a DNA sequence assembly program, Genome Res. 9 (9) (1999) 868–877, doi:10.1101/gr.9.9.868.
[10] B. Li, C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinform. 12 (2011) 323, doi:10.1186/1471-2105-12-323.
[11] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (4) (2012) 357–359, doi:10.1038/nmeth.1923.
[12] G. Parra, K. Bradnam, I. Korf, CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes, Bioinformatics 23 (2007) 1061–1067, doi:10.1093/bioinformatics/btm071.
[13] R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smirnov, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, D.A. Natale, The COG database: an updated version includes eukaryotes, BMC Bioinform. 4 (2003) 41, doi:10.1186/1471-2105-4-41.
[14] F.A. Simao, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (19) (2015) 3210–3212, doi:10.1093/bioinformatics/btv351.
[15] R.M. Waterhouse, M. Seppey, F.A. Simao, M. Manni, P. Ioannidis, G. Klioutchnikov, E.V. Kriventseva, E.M. Zdobnov, BUSCO applications from quality assessments to gene prediction and phylogenomics, Mol. Biol. Evol. 35 (3) (2017) 543–548, doi:10.1093/molbev/msx319.
[16] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410, doi:10.1016/S0022-2836(05)80360-2.
[17] K. Katoh, J. Rozewicki, K.D. Yamada, MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization, Brief Bioinform. 20 (4) (2019) 1160–1166, doi:10.1093/bib/bbx108.
[18] J. Castresana, Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis, Mol. Biol. Evol. 17 (4) (2000) 540–552, doi:10.1093/oxfordjournals.molbev.a026334.
[19] M. Gouy, S. Guindon, O. Gascuel, SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building, Mol. Biol. Evol. 27 (2) (2010) 221–224, doi:10.1093/molbev/msp259.
[20] V. Lefort, J.E. Longueville, O. Gascuel, SMS: smart model selection in PhyML, Mol. Biol. Evol. 34 (9) (2017) 2422–2424, doi:10.1093/molbev/msx149.
[21] S. Guindon, J.F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0, Syst. Biol. 59 (3) (2010) 307–321, doi:10.1093/sysbio/syq010.