Virus Genes (2012) 45:316–326
DOI 10.1007/s11262-012-0762-5

Phylogenetic analysis of *Ostreococcus* virus sequences from the Patagonian Coast

Julieta M. Manrique · Andrea Y. Calvo · Leandro R. Jones

Received: 24 March 2012 / Accepted: 11 May 2012 / Published online: 7 June 2012

Abstract A phylogenetic analysis of new *Ostreococcus* virus (OV) sequences from the Patagonian Coast, Argentina, and homologous sequences from public databases was performed. This analysis showed that the Patagonian sequences represented a divergent viral clade and that the rest of OV sequences analyzed here were clustered into six additional phylogenetic groups. Analyses of 18S gene libraries supported a close relationship of the Patagonian *Ostreococcus* host with clade A sequences described elsewhere, corroborating previous studies indicating that clade A strains are ubiquitous. Besides the Patagonian OV sequences, several phylogenetic groupings were linked to particular geographic locations, suggesting a role for allopatric cladogenesis in viral diversification. However, and in agreement with previous observations, other viral lineages included sequences with diverse geographic origins. These findings, together with analyses of ancestral trait trajectories performed here, are consistent with an evolutionary dynamics in which geographical isolation has a role in OV diversification but can be followed by rapid dispersion to remote places.

Keywords Prasinovirus · Phycodnavirus · *Ostreococcus* virus · Patagonia

Electronic supplementary material The online version of this article (doi:10.1007/s11262-012-0762-5) contains supplementary material, which is available to authorized users.

J. M. Manrique · A. Y. Calvo · L. R. Jones (✉)
Division of Molecular Biology, Estación de Fotobiología Playa Unión, CC 15 (9103), Playa Unión, Rawson, Chubut, Argentina
e-mail: ljones@conicet.gov.ar

Introduction

Until relatively recent times, environmental viruses were not given the attention they deserve. Now it is known that these microorganisms are extremely abundant and diverse, and that they are responsible for important ecological functions [1–4]. The study of biogeographic patterns of aquatic viruses has been recently boosted by the use of culture-independent molecular techniques [5–11]. Understanding the speciation mechanisms of microorganisms can be relevant not only for basic research but also for controlling emerging pathogens, forensic analysis, bioremediation and designing conservation strategies for endemic microorganisms [12]. The abundance of free-living forms suggests that most microorganisms can readily experience a global dispersal and exert rampant invasions. This line of thinking has lead to the idea that "everything is everywhere, but, the environment selects," which postulates that all microbial forms have a worldwide distribution, but most of these forms are present at negligible frequencies in particular environmental settings [13, 14]. The available data are controversial regarding whether this paradigm is correct, because the enormous diversity exhibited by viruses and other microorganisms suggests that allopatric speciation has a significant role in microbial diversification [10, 11].

The family *Phycodnaviridae* comprises six virus genera (*Chlorovirus*, *Prasinovirus*, *Prymnesiovirus*, *Phaeovirus*, *Coccolithovirus* and *Raphidovirus*) of icosahedral, dsDNA viruses that infect marine and fresh water eukaryotic algae [15–19]. The members of the genus *Prasinovirus* [17, 20] infect members of the Mamiellophyceae algal Class, which includes the smallest known photosynthetic eukaryotes [21, 22]. Photosynthetic picoplankton (<3 μm) is responsible for the majority of primary production in the

oceans [23]. In oligotrophic oceanic ecosystems, this group is dominated by cyanobacteria of the genera *Synechococcus* and *Prochlorococcus*. However, photosynthetic picoeukaryotes of the genera *Bathycoccus*, *Micromonas* and *Ostreococcus* are prevalent primary producers in estuarine waters [24, 25]. As virus-mediated cellular lysis is a major cause of cell mortality, the study of viruses that infect picoeukaryotic phytoplankton is relevant for understanding the dynamics of these important primary producers [1, 26].

Viruses infecting *Ostreococcus* sp. (OVs) can be present in coastal and open sea waters. OV sequences have been described from different *Ostreococcus* strains and relatively remote geographic locations [17, 20, 27–32]. Herein, we describe a phylogenetic analysis of new prasinovirus sequences amplified from sea waters from the Patagonian coast and homologous sequences from around the World. The information derived from these studies was combined with geographical and host information to shed light on the evolution of virus–host associations and phylogeography of OVs.

Materials and methods

Sampling

Three surface water samples (10 L) were taken on January 6, 12 and 17 of 2010, in the proximities of the mouth of the Chubut River (43.34657°S, 65.01662°W; 43.33982°S, 65.01260°W and 43.34003°S, 65.02472°W, respectively). Sampling was performed from a pneumatic boat using an acid-cleaned opaque carboy-tank, during the high tide (seawater condition). All samples were immediately transported to the laboratory and processed as described below.

Co-immobilization of virus–host assemblages

To enrich the samples in the fraction of the virome corresponding to *Ostreococcus* viruses, 2 L from each of the three water samples was successively filtered through a series of filters of decreasing pore size (20, 10, 5, 1.2 and 0.22 μm). The filters were stored at −80 °C until used for nucleic acids extraction as described in the next section. The approach of amplifying viral sequences present in host cells is novel, as most previous studies of viruses in this group rather focus on free virions. One of the advantages of our approach is that most of the viral sequences obtained are likely to be replicating, though we cannot exclude the possibility that some sequences could correspond to free virions that were adhering to cell surfaces.

Nucleic acids extraction and PCR amplification

Total DNA retained in the filters was extracted using the following modification of Doyle & Doyle's protocol [33]: The filters were incubated at 60 °C for 30 min with 720 μL of pre-heated CTAB buffer (2 % [w/v] CTAB Sigma, 1.4-M NaCl, 0.2% [v/v] β-mercaptoethanol, 20-mM EDTA, 100-mM Tris–HCl pH 8.0). After treatment with chloroform/isoamyl alcohol, the suspension was treated with RNase A (Sigma-Aldrich) at a final concentration of 10 μg/mL for 1 h at 37 °C. The RNase was removed with chloroform, and nucleic acids were precipitated with isopropanol. The sample was centrifuged at 20,000×*g* for 30 min at 4 °C, and the pellet was washed with 70 % ethanol. The DNA pellet was dried and resuspended in ultrapure, DNase-free water (Invitrogen). The obtained amount of DNA was estimated by densitometric analysis against a standard curve (High DNA Mass Ladder, Invitrogen) using the ImageJ software [34].

The DNAs obtained as described earlier were used, separately, as templates for amplifying 18S sequences and a portion of the viral polymerase gene. The 18S gene analyses were performed by amplifying a region of the gene of ~1,517 bp. The universal primers used to this aim were UNI_17F and UNI_1534R [35]. The amplification of the viral polymerase gene was performed using the algal-virus-specific AVS primer set [17, 36]. To verify that the correct target was amplified, a second-round nested PCR was performed with AVS-1 and POL primers [36, 37].

All the reactions were performed using 50-μL reaction mixtures, 1 U of AccuPrime™ TaqDNA Polymerase High Fidelity (Invitrogen), 5 μL of 10× AccuPrime™ Buffer II (Invitrogen) and a primer concentration of 1 μM for UNI_17F/UNI_1534R or 1.2 μM for the AVS-1/AVS-2 and AVS-1/POL. An amount of 50–100 ng of template DNA was used for UNI_17F/UNI_1534R and AVS-1/AVS-2 amplifications, whereas 1 μL from the gel-purified band was used for AVS-1/POL amplification. For all the primer pairs used in this study, the optimal annealing temperature was fine-tuned by gradient PCR. The optimal annealing conditions were 62.2, 45.3 and 51.3 °C for the 18S, AVS and the AVS-1/POL primers sets, respectively. The thermal cycling was performed with an initial denaturation step (94 °C for 60 s) followed by 35 amplification cycles (94 °C for 30 s, optimized annealing temperature for 30 s, and 1 min/kb extension at 68 °C). All the reactions were carried out in a BioRad My Cycler thermal cycler (Bio-Rad Laboratory, Inc.). Both the 18S and the polymerase PCRs from each filter were performed in triplicate and the obtained products were pooled to avoid PCR biases. These products were purified by the QIAquick PCR Purification Kit (QIAGEN), and then quantitated with a Nanovue Plus spectrofotometer (GE Health care).

Molecular cloning and sequencing

The amplicons obtained as described earlier were cloned into a pGem vector (pGEM®-T Easy vectorSystem II kit, Promega), using *Escherichia coli* strain TOP 10 (Invitrogen) to obtain three 18S and three polymerase gene libraries. Recombinant plasmids were partially purified as described elsewhere [38] and used as template for insert PCR amplification. The plasmid inserts were amplified using pGEM®-T-specific primers T7 and SP6. The corresponding PCR mixtures were set up with 1 μL of the plasmid preparation as template, 1 U of AccuPrime™ TaqDNA Polymerase High Fidelity (Invitrogen), 5 μL of 10× AccuPrime™ Buffer I (Invitrogen), 1 μL of dNTPs (10 mM each), 2.5 μL of each primer (20 μM) and ultrapure, DNase-free water (Invitrogen) to a final volume of 50 μL. The amplification conditions consisted of an initial denaturation (94 °C for 60 s) followed by 30 cycles of amplification (94 °C for 30 s, 53 °C for 30 s and 68 °C for 1 min/kb extension). The amplified inserts were quantitated and sequenced in both directions using specific primers. The presence of chimeras was checked using the de novo mode of the program UCHIME [39]. Both the 18S and the polymerase haplotypes were randomly distributed among the libraries (Pearson's chi-square test), indicating that no biases in the haplotypes' frequencies were introduced by the experimental procedures. The 118 sequences described here have GenBank accession numbers JQ691949–JQ692067.

Datasets

To construct a comprehensive dataset, the Patagonian sequences were combined with previously published OV sequences (Table 1). Sequences from *Bathycoccus* and *Micromonas* viruses were used as outgroup. For the hosts' dataset, sequences from several members of the five previously described *Ostreococcus* clades [21, 40] (accession numbers AY425307, AY425308, AY425310, AY425311, AY425313, AB058376, NC_014437, GQ426331, GQ426335, GQ426336–GQ426338, GQ426340–GQ426347, AY329635, Y15814 and AY329636) and outgroup sequences from *Bathycoccus* (accession numbers FN562453, AY425315 and AY425314) and *Micromonas* (accession numbers HM191693, DQ025753, FN562452, AY954993–AY955012, AB183589 and AJ010408) species were included. Both datasets were aligned using the program MAFFT, with default *op* and *ep* and the weighted sum-of-pairs score and consistency score obtained from local alignments [41]. Sequence alignments were visually inspected using *Genetic Data Environment* [42, 43] for MacOS X (MacGDE 2.4). After that, a codon-wise alignment was obtained using the Los Alamos Codon Alignment tool, which is available at http://www.hiv.lanl.gov/content/sequence/CodonAlign/codonalign.html. The numbers of synonymous and non-synonymous substitutions [44, 45] were obtained with the program SNAP [46].

Haplotype network analyses

Statistical parsimony network analyses [47] were performed with the program TCS [48], using a connection probability of 95 % [49, 50].

Similarity analyses

Sequence similarities were obtained as described elsewhere [51, 52]. Shortly, for a given sequence pair a and b, the number of nucleotide or amino acid substitutions (D) was obtained as follows:

$$D_{ab} = \sum_{i=1}^{P} f(a_i, b_i),$$

where P is the number of aligned positions and $f(a_i,b_i)$ is obtained by the following equation:

$$f(a_i, b_i) = \begin{cases} 0, & \text{if } a_i = b_i \\ 1, & \text{if } a_i \neq b_i \end{cases}$$

The inter-cluster comparisons were averaged by the number of sequence comparisons performed in each case [51, 52].

Phylogenetic analyses

Phylogenies were inferred using parsimony, maximum likelihood and Bayesian techniques. The parsimony analyses were performed with the program TNT [53], using the approaches described in refs. [54] and [55]. The maximum likelihood trees were obtained by PhyML 3.0 [56], under evolutionary models inferred with MrAIC.pl [57]. PhyML was set to estimate model parameters during the searches. The Bayesian phylogenetic analyses were performed with MrBayes 3.2.1 [58, 59]. Twelve Monte Carlo chains (MCMC) were run for 10E7 generations, sampling every 1,000 generations. Posterior probabilities were calculated and reported on a 50 % majority rule consensus tree of the post-burnin sample. Genetic differentiation indices were obtained as described elsewhere [60]. Here, we also implemented an equivalent index l_i/L, where l_i is the length of a test split i in the ML tree and L is the median of all the branch lengths in the ML tree. The median is used for scaling because the branch length distributions from trees with divergent lineages are asymmetrical. Being scaled in this way, these indices' units are proportional to the amount of diversity contributed by the tested branch. Phylogenetic trees were visualized and drawn using the program Dendroscope [61].

Table 1 Viral sequences analyzed in this work

DBA[a]	Strain	Network[b]	Genotype[c]	Origin[d]	Clade[e]
NC_014765	BpV1	–	–	MS	–
HM004430	BpV2	–	–	MS	–
NC_014767	MpV1	–	–	MS	–
MPU32975	SP1	–	–	NAO	–
MPU32976	SP2	–	–	NAO	–
MPU32982	PL1	–	–	NPO	–
MPU32981	SG1	–	–	NPO	–
MPU32980	PB8	–	–	NAO	–
MPU32979	PB7	–	–	NAO	–
MPU32978	PB6	–	–	NAO	–
MPU32977	GM1	–	–	NPO	–
NC_014766	OlV1	1	2	MS	A
GQ412099	OlV158	1	2	MS	A
GQ412100	OlV164	1	2	MS	A
GQ412091	OlV462	2	2	NAO	A
GQ412095	OlV467	2	2	NAO	A
GQ412092	OlV464	2	2	NAO	A
GQ412093	OlV465	2	2	NAO	A
GQ412094	OlV466	2	2	NAO	A
GQ412098	OlV458	2	2	NAO	A
JQ691969	OsVPU[f]	3	1	SAO	A
JQ691949	OsVPU	3	1	SAO	A
JQ691960	OsVPU	3	1	SAO	A
JQ691951	OsVPU	3	1	SAO	A
JQ692032	OsVPU	3	1	SAO	A
JQ691952	OsVPU	3	1	SAO	A
JQ691996	OsVPU	3	1	SAO	A
GQ412082	OlV349	4	5	EC	A
GQ412083	OlV350	4	5	EC	A
GQ412090	OlV470	4	5	EC	A
GQ412089	OlV468	4	5	EC	A
GQ412088	OlV402	4	5	EC	A
GQ412084	OlV359	4	5	SPO	A
NC_014789	Otv-2	4	5	EC	B
JCVI_READ_1092963530480	–	4	5	NAO	–
GQ412085	OlV360	5	5	SPO	A
GQ412096	OlV536	5	5	MS	A
GQ412097	OlV537	5	5	MS	A
GQ412101	OlV155	5	5	MS	A
GQ412086	OlV364	6	5	SPO	A
GQ412087	OlV368	6	5	SPO	A
FJ267509	OTV102	7	7	MS	C
FJ267510	OTV113	7	7	MS	C
FJ267504	OTV3	7	7	MS	C
FJ267498	OTV23	7	7	MS	C
FJ267512	OTV126	7	7	MS	C
FJ267496	OTV21	7	7	MS	C
FJ267505	OTV29	7	7	MS	C

Table 1 continued

DBA[a]	Strain	Network[b]	Genotype[c]	Origin[d]	Clade[e]
FJ267508	OTV78	7	7	MS	C
FJ267506	OTV52	7	7	MS	C
FJ267507	OTV72	7	7	MS	C
NC_010191	OtV5	7	7	MS	C
NC_013288	OTV-1	7	7	EC	C
FJ267511	OTV121	7	7	MS	C
FJ267497	OTV22	7	7	MS	C
FJ267502	OTV64	8	6	MS	D
FJ267499	OTV66	8	6	MS	D
FJ267500	OTV67	8	6	MS	D
FJ267501	OTV63	9	4	MS	D
FJ267503	OTV65	9	4	MS	D
AF405581	BSA99_5	–	3	NPO	–
EU889370	KBvp_17	–	3	NPO	–
JCVI_READ_1092955067637	–	–	6	NAO	–
JCVI_READ_1091140189807	–	–	–	NAO	–
JCVI_READ_1092246500722	–	–	3	NAO	–

[a] Database accession number

[b] Genetic groupings identified by statistical parsimony analysis

[c] Phylogenetic groupings identified by phylogenetic species recognition analyses

[d] Geographic origin of the viral strain or sequence. *MS* Mediterranean Sea; *SAO* South Atlantic Ocean; *NAO* North Atlantic Ocean; *NPO* North Pacific Ocean; *GOM* Gulf of Mexico; *EC* English Channel; *SPO* South Pacific Ocean

[e] *Ostreococcus* sp. clade

[f] OsVPU: *Ostreococcus* Virus Playa Unión. Only one sequence from each haplotype was included in phylogenetic analyses. For the full list of GenBank accession numbers see "Materials and Methods" section

Statistical analyses

All statistical tests were performed with the R statistical package [62]. Non-phylogenetic tests were made using Fisher's exact tests, with p values obtained by Monte Carlo simulation.

Analyses of ancestral trait trajectories

Ancestral character state estimations and analyses of the branch scaling *kappa* (κ) parameter [63, 64] were done with the program BayesTraits, which is available for download at www.evolution.rdg.ac.uk. To cope with phylogenetic uncertainty, a sample of 2,000 trees was obtained from the stationary states of four independent MCMC chains using BayesPhylogenies (available for download at www.evolution.rdg.ac.uk). The first 20,000 points of the MCMC chains were discarded as burn-in. After that, trees were sampled at intervals of 20,000 trees to ensure that the sampled trees were statistically independent from each other. One-parameter models were used for the analyses of both host and geographical ranges. Uninformative priors were used, with intervals based on parameter values obtained in preliminary ML analyses. The likelihood ratio (LR) test was used to check for statistical significance in ML comparisons [63].

Results

The co-immobilization of virus–host complexes was corroborated by the amplification of DNA fragments of the expected size from all the 0.22-μm filters. The corresponding viral sequences displayed high similarities to OV sequences available in public databases and were highly similar to each other, indicating that the viral population constituted a single viral species. Eighty-six percent of the sequences amplified with the 18S primers were highly similar to *Ostreococcus* sequences present at databases. These sequences also were highly similar to each other, indicating that they corresponded to a single *Ostreococcus* sp. host. The rest of 18S sequences corresponded to putative *Scuticociliatia* sp. (Eukaryota, Alveolata, Ciliophora), *Girodinium* sp. (Eukaryota, Alveolata, Dinophyceae, Gymnodiniales, Gymnodiniaceae), *Cafeteria* sp. (Eukaryota, stramenopiles, Bicosoecida, Cafeteriaceae) and *Alexandrium tamarense*

(Eukaryota, Alveolata, Dinophyceae, Gonyaulacales, Gonyaulacaceae) sequences.

The OV sequences from Patagonia were combined with viral sequences from elsewhere (Table 1). This dataset consisted of 478 nucleotide positions, of which 238 presented polymorphisms and 217 harbored informative variability. The presence of polymorphisms was mostly because of synonymous substitutions; however, some positions also presented non-synonymous substitutions (Fig. 1). Statistical parsimony analyses of these sequences resulted in ten networks and eight singletons. The outgroup sequences clustered separately from the ingroup ones. The two sequences from *Bathycoccus* viruses, as well as one of the *Micromonas* virus sequences (NC 014767), constituted three separate singletons, whereas the rest of sequences from *Micromonas* viruses were grouped into a single network (not shown). Five of the OV sequences could not be connected to any of the other haplotypes. The rest of OV sequences constituted nine networks, with the Patagonian sequences clustered into a single group (Table 1).

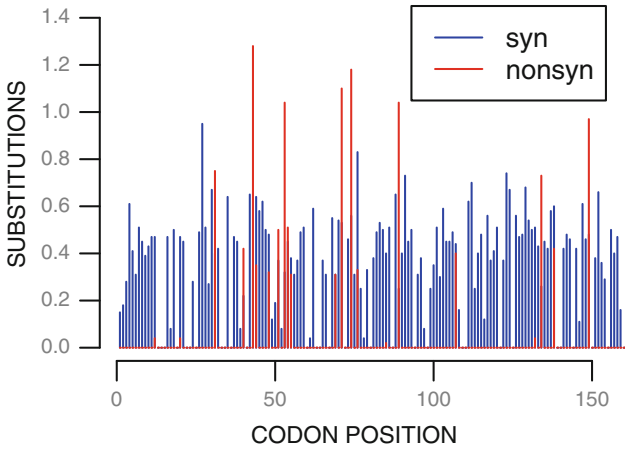

Fig. 1 Mean numbers of synonymous and non-synonymous substitutions in codon-wise paired comparisons among the OV sequences studied

Networks 4, 5 and 6 were relatively similar to each other, suggesting that these sequences could correspond to a single viral lineage (Table 2). In fact, when the sequences from clusters 4, 5 and 6 were pooled together and the mean pairwise sequence similarity was calculated, the obtained value (11.83) was similar to the intra-network similarities. Networks 1 and 2 also displayed a reciprocal similarity that was higher than the average one (Table 2).

Phylogenetic analyses were quite congruent with the statistical parsimony and similarity ones, with all the viral networks but network 6 corresponding to monophyletic groups (Fig. 2; Table 3). These analyses also indicated that three singleton sequences (EU889370, AF405581 and JCVI_READ_1092246500722) constituted a further viral clade, and that the sequence JCVI_READ_1092955067637 was related to the viruses clustered into network 8. Networks 4, 5 and 6, as well as networks 1 and 2, presented a relatively weak genetic differentiation, and some of them were poorly supported in phylogenetic analyses (Table 3). Based on these observations, we grouped the OV sequences studied here into seven genotypes, as indicated in Fig. 2 and Table 1. Each of these genotypes was strongly supported by bootstrap values and posterior probabilities, and highly divergent in relation to the rest of OV sequences (Table 4).

To place the *Ostreococcus* sequences within a specific clade, our data were combined with other Mamiellophyceae sequences to generate a dataset [21, 40]. The obtained dataset presented 1,623 nucleotide positions, of which 194 were variable and 154 were informative. The statistical parsimony analysis clustered the Patagonian sequences with previously described clade A sequences from elsewhere (Fig. 3). Phylogenetic analyses confirmed that the Patagonian sequences corresponded to an *Ostreococcus* sp., and also supported a close relationship with other Clade A strains (Figs. S3–S5). The presence of a clade A strain in Patagonia corroborates previous results indicating that clade A strains are ubiquitous [27].

Table 2 Mean number of inter-network nucleotide (lower triangle) or amino acid (upper triangle) substitutions

Network	1	2	3	4	5	6	7	8	9
1	*0.00*	0.94	8.86	5.76	5.53	5.14	10.04	6.90	6.85
2	21.63	*1.00*	9.06	6.85	6.72	6.46	11.14	7.89	7.38
3	69.54	72.46	*1.71*	10.10	9.93	9.60	10.32	9.50	5.86
4	78.12	89.67	94.14	*1.60*	0.00	0.00	8.21	8.32	8.47
5	74.07	86.88	86.72	21.15	*4.66*	0.00	8.14	8.00	8.00
6	71.14	85.07	84.53	12.11	11.11	*0.00*	8.00	7.42	7.20
7	85.67	90.25	96.72	72.65	70.54	69.17	*10.40*	6.69	7.93
8	79.20	86.21	88.81	78.96	79.92	74.00	60.348	*15*	6.57
9	79.71	84.92	81.86	85.05	86.22	77.60	80.20	65.42	*0.00*

Intra-network nucleotide distances are given in the *diagonal*

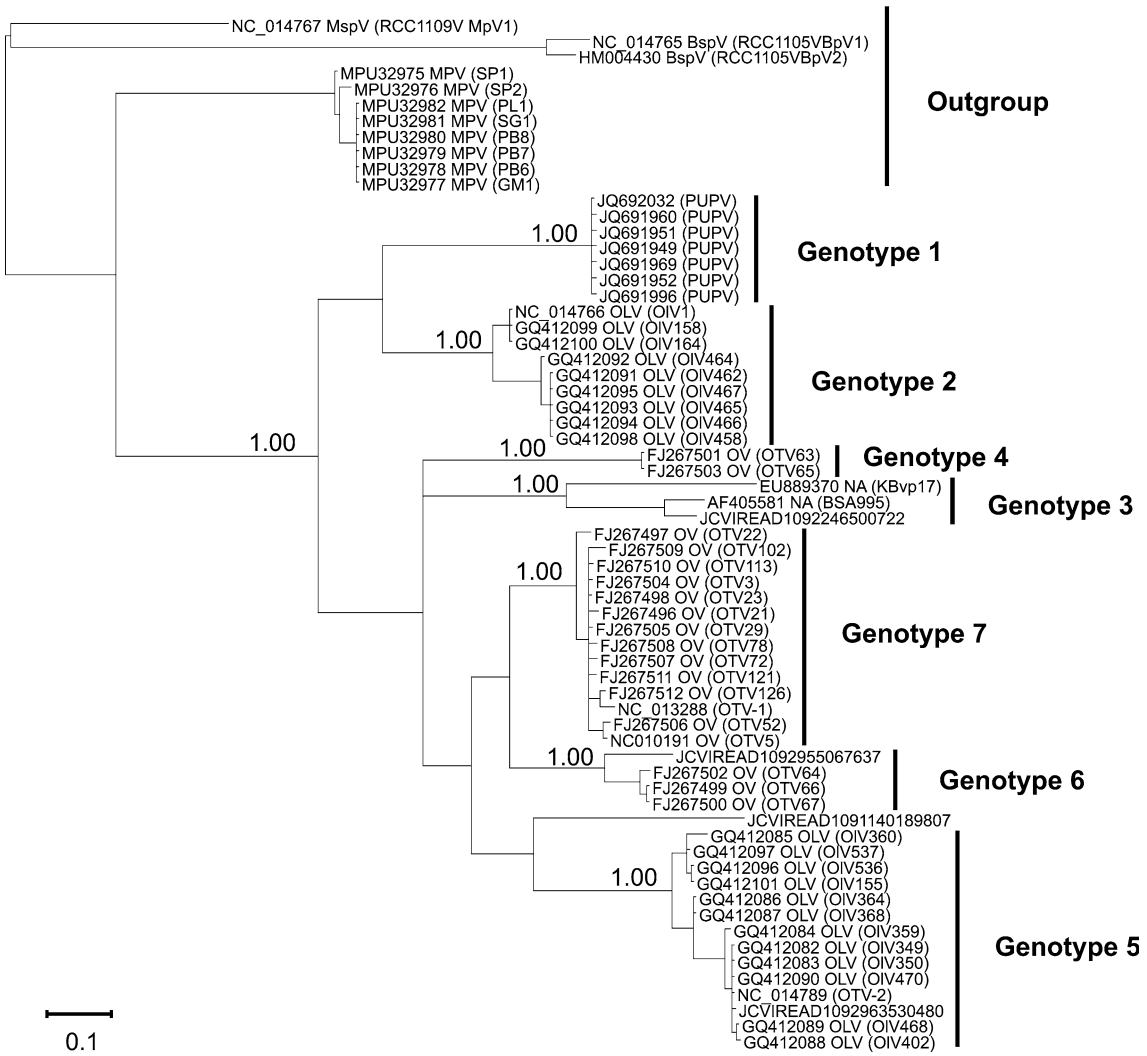

Fig. 2 Bayesian phylogenetic tree of the OV strains studied here. Numbers close to branches correspond to posterior probabilities. Branch lengths are proportional to the number of nucleotide substitutions (the scale bar units are substitutions per aligned position). Equivalent results were obtained by parsimony and maximum likelihood analyses (Figs. S1, S2). Strain names are given in *parentheses*. *MspV Micromonas* sp. virus; *BspV Bathycoccus* sp. virus; *MPV Micromonas pusilla* virus; *OLV Ostreococcus lucimarinus* virus; *OV Ostreococcus* virus; *NA* not available

Table 3 Genetic differentiation indices and phylogenetic supports of the statistical parsimony networks

Network	1	2	3	4	5	6	7	8	9
s_i/S	5	9	35	8	2	1	15	8	37
b_i/B	2.7	8	34.4	5	2.3	–[d]	10.9	5.7	36.1
l_i/L	4.7	18	80	11.4	3.3	0.009	20.9	13.3	82.1
PB[a]	94	100	100	99	54	56	100	94	100
MLB[b]	94	95	100	94	38	100	90	66	100
P[c]	0.78	1.00	1.00	1.00	0.65	–	1.00	0.90	1.00

[a] Bootstrap support in parsimony analysis

[b] Bootstrap support in ML analysis

[c] Posterior probability

[d] Poly- or paraphyletic

Table 4 Genetic differentiation indices and phylogenetic supports of the genotypes identified here

Genotype	1	2	3	4	5	6	7
s_i/S	35	21	19	37	22	13	15
b_i/B	34.4	18.1	23.6	36.1	22.8	15.6	10.9
l_i/L	80.2	32.3	50.3	82.1	48.4	35.1	20.9
PB[a]	100	98	88	100	100	92	100
MLB[b]	100	90	97	100	97	83	90
P[c]	1.00	1.00	1.00	1.00	1.00	1.00	1.00

[a] Bootstrap support in parsimony analysis

[b] Bootstrap support in ML analysis

[c] Posterior probability

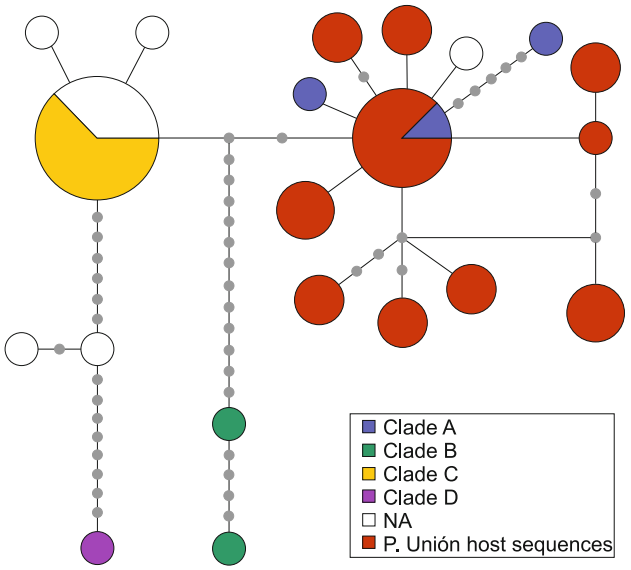

Fig. 3 Haplotype network for the *Ostreococcus* 18S gene under the 95 % parsimony criterion. The *large colored circles* correspond to groupings of observed haplotypes, with the circles' radiuses proportional to the number of accrued sequences. The *smaller gray dots* indicate missing haplotypes. *NA* not available, *P* Unión host sequences: host sequences from the Patagonian coast

In agreement with previous studies [27], many viral sequences from remote locations were highly related to each other, with the outstanding example of genotype 5, which was highly diverse in terms of geographic origin (Table 1). However, non-phylogenetic statistical analyses indicated that geographical ranges were associated to genetic grouping (Table 1; $p = 4E{-}4$). In addition, the analyses of the branch scaling parameter κ indicated a gradual phylogeny coupled mode of evolution for this trait (Fig. 4) [63, 65], and character state estimations suggested narrow ancestral geographical ranges for genotypes 1, 4 and 7 (Fig. S6) and for the majority of statistical parsimony networks (Fig. S7). Approximately half of the points sampled from the posterior distribution of trees indicated, for the geographical range, a significantly smaller than one κ value (Fig. 4), suggesting that geographical range variance reached a maximum while the overall divergence was still progressing. As expected [17, 27], not all the OV groupings identified here were present in all the previously described *Ostreococcus* clades (Table 1), and host ranges were associated with genetic groupings (Table 1; $p = 1E{-}5$). Furthermore, the analysis of the k parameter supported a cophylogenetic structure in this virus–host system (Fig. 4), and Bayesian character state estimates indicated single ancestral hosts for the majority of viral lineages (Fig. S8). However, *Ostreococcus* strains from clades A and D harbored multiple viral genotypes, and an analysis made with the most recent common ancestor (MRCA) method [64] supported clade A and D *Ostreococcus* hosts for the MRCAs of genotypes 1, 2 and 5, and 4 and 6, respectively (Fig. S9).

Discussion

The analyses described here show that the studied OV sequences from Patagonia correspond to a divergent viral clade and that the rest of OV sequences can be classified into six further phylogenetic groupings (Fig. 2). One sequence (JCVI READ 1091140189807) failed to cluster with other sequences and thus could represent an eighth genotype. The B-family DNA polymerase is highly conserved at the amino acid level and less conserved at the nucleotide one, allowing for accurate phylogenetic inference at different taxonomic categories [7, 8, 17, 27, 36, 37, 66, 67]. However, a point that deserves further consideration is whether the genetic patterns observed here using DNA polymerase data can be extrapolated to the rest of the viral genome. Comparisons between the available OV genomes demonstrated a high degree of colinearity. However, exclusive coding sequences also have been identified in all the viral genomes studied so far [29]. The

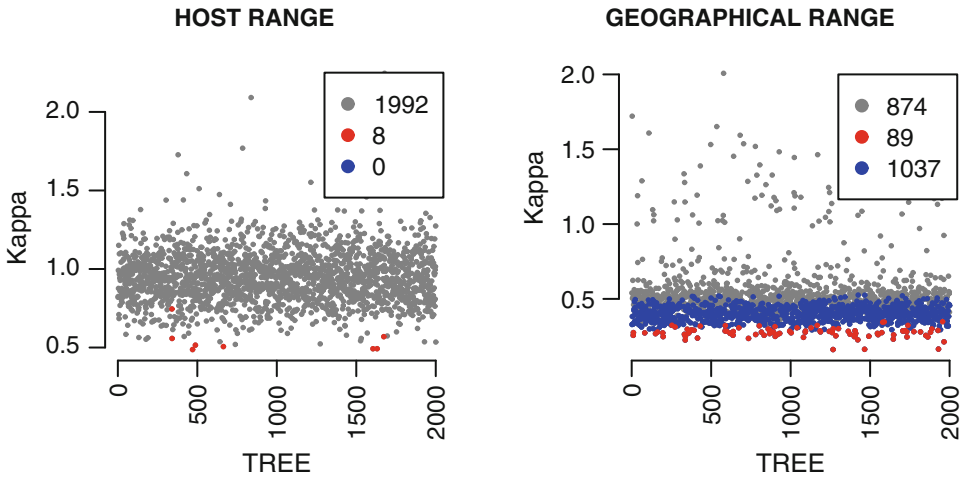

Fig. 4 Maximum likelihood estimates of the Kappa (κ) parameters and evaluation of the nulls that $\kappa = 0$ and $\kappa = 1$ against $\kappa > 0$ and $\kappa \neq 1$, respectively. The estimations were performed on a sample of 2,000 points taken from the posterior distribution of trees as described in the "Materials and methods" section. The *red points* represent trees for which κ was not significantly larger than zero. The trees for which κ was significantly smaller (or larger) than 1 are indicated in *blue*

genome of OV strain OtV-2, which clustered into genotype 5, encodes a cytochrome b_5, an RNA polymerase sigma factor, a high-affinity phosphate transporter, a putative deoxycytidylate deaminase, a putative 6-phosphogluconate dehydrogenase, a putative 2OG-Fe(II) oxygenase and a putative tail fiber assembly protein that are not present in the genomes of strains OtV-1 and OtV-5, which belonged to genotype 7 [29]. On the other hand, the genomes of OtV-1 and OtV-5 also encode proteins that are absent from the OtV-2 one [29]. These data suggest that inter-genotype genomic diversity might be relevant for the study of OVs.

Previous studies have shown that prasinoviruses from *Ostreococcus*, *Bathycoccus* and *Micromonas* species are monophyletic and host specific [17, 20, 28], which is consistent with the idea that nucleocytoplasmic large DNA viruses (NCLDV) have evolved following the radiation of eukaryotes [68]. In agreement with these studies, we observed that not all the OV groupings identified here were present in all the previously described *Ostreococcus* clades, and our evolutionary analyses indicated single hosts for the majority of genotypes' and networks' MRCAs (Figs. S8, S9). Nonetheless, the viruses clustered into genotypes 1, 2 and 5 come from clade A *Ostreococcus* strains, with the exception of a single strain (OtV-2), which was obtained from a clade B one. Likewise, the sequences grouped into genotypes 4 and 6 come from viruses infecting clade D *Ostreococcus* strains. These last observations support that the evolution of these viral lineages was decoupled from the evolution of their host. The inclusion of a single OV strain from a clade B *Ostreococcus* into genotype 5 (OtV-2) suggests the occurrence of horizontal transfer. However, this should be confirmed by identifying a larger amount of viral strains with these characteristics.

Our evolutionary analyses indicated narrow ancestral geographical ranges for three viral genotypes and eight statistical parsimony networks (Figs. S6, S7). Remarkably, besides the Patagonian sequences, several viral lineages infecting clade A *Ostreococcus* strains were associated to particular geographical regions: Networks 1 and 2 included sequences from the Mediterranean Sea and the North Atlantic Ocean, respectively, the two network 6 sequences were from the South Pacific Ocean and network 5 MRCA had a restricted ancestral distribution (Table 1, Fig. S7). Host dispersal is known to be a driver of viral diversification [69, 70] and implies a leptokurtic form of dissemination in which long-distance spread is achieved by small proportions of the source population [71]. This determines the establishment of isolated geographical groups within which genetic variation is low in comparison to inter-group variability. This agrees with our findings, as we observed clades that included relatively similar sequences, that were separated from each other by relatively longer branches (Fig. 2; Tables 3, 4). On the other hand, and in agreement with previous studies [27], we observed no association between phylogeny and geographical patterns for genotypes 2, 3, 5 and 6 and for network 4. In addition, the prevalence of values smaller than one for the branch-scaling parameter κ (Fig. 4) indicates that geographical range diversification reached a maximum while the overall divergence was still progressing, which is consistent with a scenario in which geographical dispersion has followed diversification rapidly [63, 65]. Altogether, these findings suggest an evolutionary dynamics in which geographical isolation participates in viral diversification but is followed by rapid dispersion to remote geographic locations, which could explain why some viral lineages were geographically

structured and some others were not. Given that *Ostreococcus* strains from different clades present adaptations to different light conditions [40], another plausible driver of viral diversification could be the acquisition of selective advantages in particular light niches [29]. In agreement with this idea, geographically close but contrasting environments can share few variant viral polymerase gene sequences [66]. This could be the case of genotype 4 and network 8, which grouped viral strains from relatively nearby locations, all of which were obtained from clade D *Ostreococcus* strains, which are capable of growing at a wide range of light intensities [40].

Our analyses show that OVs constitute an interesting model for the study of microbial biogeography. Although a comprehensive OV dataset was used in this study, we think that further studies will be needed to fully understand OV phylodynamics. For example, all but one of the strains clustered into genotype 7 were of Mediterranean Sea origin. This suggests that sampling more sequences from other genotypes, such as, for example, genotype 4, which in our dataset was represented by two strains, could provide further evidence of dispersal. Likewise, studies integrating information on the environmental conditions regnant at the locations from which viruses were isolated with data on these viruses' genetic background will help to decipher the role of adaptation in OV diversification.

Acknowledgments Continuous support from National Council for Scientific and Technical Research (CONICET, Argentina) is deeply appreciated. Support from the Agencia Nacional de Promoción Científica y Técnica is also acknowledged. We want to thank Pablo Goloboff for TNT assistance and guidance in phylogenetic analyses. We also are in debt with Andrew Meade for his suggestions and assistance with Bayesian analyses and the use of BayesTraits and BayesPhylogenies.

References

1. C.A. Suttle, Nat. Rev. **5**, 801–812 (2007)
2. K.E. Wommack, R.R. Colwell, Microbiol. Mol. Biol. Rev. **64**, 69–114 (2000)
3. M.J. Allen, W.H. Wilson, Curr. Opin. Microbiol. **11**, 226–232 (2008)
4. M. Filippini, M. Middelboe, FEMS Microbiol. Ecol. **60**, 397–410 (2007)
5. F.E. Angly, B. Felts, M. Breitbart, P. Salamon, R.A. Edwards, C. Carlson, A.M. Chan, M. Haynes, S. Kelley, H. Liu, J.M. Mahaffy, J.E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C.A. Suttle, F. Rohwer, PLoS Biol. **4**, e368 (2006)
6. J.L. Clasen, C.A. Suttle, Appl. Environ. Microbiol. **75**, 991–997 (2009)
7. M.V. Gimenes, P.M. Zanotto, C.A. Suttle, H.B. da Cunha, D.U. Mehnert, ISME J. 1–11 (2011)
8. S. Short, O. Rusanova, M.A. Staniewski, Aquat. Microb. Ecol. **63**, 61–67 (2011)
9. S.J. Williamson, D.B. Rusch, S. Yooseph, A.L. Halpern, K.B. Heidelberg, J.I. Glass, C. Andrews-Pfannkoch, D. Fadrosh, C.S. Miller, G. Sutton, M. Frazier, J.C. Venter, PLoS ONE **3**, e1456 (2008)
10. M. Breitbart, B. Felts, S. Kelley, J.M. Mahaffy, J. Nulton, P. Salamon, F. Rohwer, Proc. R. Soc. B **271**, 565–574 (2004)
11. M. Breitbart, P. Salamon, B. Andresen, J.M. Mahaffy, A.M. Segall, D. Mead, F. Azam, F. Rohwer, Proc. Natl. Acad. Sci. USA **99**, 14250–14255 (2002)
12. R.J. Whitaker, Philos. Trans. R. Soc. London B **361**, 1975–1984 (2006)
13. R. de Wit, T. Bouvier, Environ. Microbiol. **8**, 755–758 (2006)
14. M.A. O'Malley, Stud. Hist. Phil. Biol. Biomed. Sci. **39**, 314–325 (2008)
15. J.L. Van Etten, M.V. Graves, D.G. Muller, W. Boland, N. Delaroque, Arch. Virol. **147**, 1479–1516 (2002)
16. W.H. Wilson, J.L. Van Etten, M.J. Allen, Curr. Top. Microbiol. Immunol. **328**, 1–42 (2009)
17. L. Bellec, N. Grimsley, H. Moreau, Y. Desdevises, Environ. Microbiol. Rep. **1**, 114–123 (2009)
18. J.B. Larsen, A. Larsen, G. Bratbak, R.A. Sandaa, Appl. Environ. Microbiol. **74**, 3048–3057 (2008)
19. D.D. Dunigan, L.A. Fitzgerald, J.L. Van Etten, Virus Res. **117**, 119–132 (2006)
20. H. Moreau, G. Piganeau, Y. Desdevises, R. Cooke, F. Derelle, N. Grimsley, J. Virol. **84**, 12555–12563 (2010)
21. L. Guillou, W. Eikrem, M.-J. Chrétiennot-Dinet, F. Le Gall, R. Massana, K. Romari, C. Perdós-Alió, D. Vaulot, Protist **155**, 193–214 (2004)
22. B. Marin, M. Melkonian, Protist **161**, 304–336 (2010)
23. F.P. Chavez, M. Messie, J.T. Pennington, Ann. Rev. Marine. Sci. **3**, 227–260 (2011)
24. I.C. Biegala, F. Not, D. Vaulot, N. Simon, Appl. Environ. Microbiol. **69**, 5519–5529 (2003)
25. P.W. Johnson, J.M. Sieburth, J. Phycol. **18**, 318–327 (1982)
26. C.P. Brussaard, J. Eukaryot. Microbiol. **51**, 125–138 (2004)
27. L. Bellec, N. Grimsley, Y. Desdevises, Appl. Environ. Microbiol. **76**, 96–101 (2010)
28. E. Derelle, C. Ferraz, M.L. Escande, S. Eychenie, R. Cooke, G. Piganeau, Y. Desdevises, L. Bellec, H. Moreau, N. Grimsley, PLoS ONE **3**, e2250 (2008)
29. K.S. Weynberg, M.J. Allen, I.C. Gilg, D.J. Scanlan, W.H. Wilson, J. Virol. **85**, 4520–4529 (2011)
30. K.D. Weynberg, M.J. Allen, K. Ashelford, D.J. Scanlan, W.H. Wilson, Environ. Microbiol. **11**, 2821–2839 (2009)
31. S.M. Short, C.A. Suttle, Appl. Environ. Microbiol. **68**, 1290–1296 (2002)
32. A.I. Culley, B.F. Asuncion, G.F. Steward, ISME J. **3**, 409–418 (2009)
33. J.J. Doyle, J.L. Doyle, Focus **12**, 13–15 (1991)
34. M.D. Abràmoff, P.J. Magalhães, S.J. Ram, Biophotonics Int. **11**, 36–42 (2004)
35. S.Y. Moon-van der Staay, R. De Wachter, D. Vaulot, Nature **409**, 607–610 (2001)
36. F. Chen, C.A. Suttle, Appl. Environ. Microbiol. **61**, 1274–1278 (1995)
37. F. Chen, C.A. Suttle, Biotechniques **18**, 609–612 (1995)
38. L.R. Jones, R. Zandomeni, E.L. Weber, J. Gen. Virol. **83**, 2161–2168 (2002)
39. R.C. Edgar, B.J. Haas, J.C. Clemente, C. Quince, R. Knight, Bioinformatics **27**, 2194–2200 (2011)
40. F. Rodríguez, E. Derelle, L. Guillou, F. Le Gall, D. Vaulot, H. Moreau, Environ. Microbiol. **7**, 853–859 (2005)
41. K. Katoh, G. Asimenos, H. Toh, Methods Mol. Biol. (New York, NY, USA) **537**, 39–64 (2009)
42. J.A. Eisen, Methods Mol. Biol. **70**, 13–38 (1997)
43. S.W. Smith, R. Overbeek, C.R. Woese, W. Gilbert, P.M. Gillevet, Comput. Appl. Biosci. **10**, 671–675 (1994)

44. M. Nei, T. Gojobori, Mol. Biol. Evol. **3**, 418–426 (1986)
45. T. Ota, M. Nei, Mol. Biol. Evol. **11**, 613–619 (1994)
46. B. Korber, in *Computational Analysis of HIV Molecular Sequences*, edited by A.G. Rodrigo, G.H. Learn (Kluwer Academic Publishers, Dordrecht, the Netherlands, 2000) p. 55
47. A.R. Templeton, K.A. Crandall, C.F. Sing, Genetics **132**, 619–633 (1992)
48. M. Clement, D. Posada, K.A. Crandall, Mol. Ecol. **9**, 1657–1659 (2000)
49. H. Chen, M. Strand, J.L. Norenburg, S. Sun, H. Kajihara, A.V. Chernyshev, S.A. Maslakova, P. Sundberg, PLoS ONE **5**, e12885 (2010)
50. M.W. Hart, J. Sunday, Biol. Lett. **3**, 509–512 (2007)
51. S.M. Rodriguez, M.D. Golemba, R.H. Campos, K. Trono, L.R. Jones, J. Gen. Virol. **90**, 2788–2797 (2009)
52. Y.C. Shin, L.R. Jones, J.M. Manrique, W. Lauer, A. Carville, K.G. Mansfield, R.C. Desrosiers, Virology **400**, 175–186 (2010)
53. P.A. Goloboff, J.S. Farris, K. Nixon, Cladistics **24**, 774–786 (2008)
54. P.A. Goloboff, Cladistics **15**, 415–428 (1999)
55. P.A. Goloboff, J.S. Farris, Cladistics **17**, S26–S34 (2001)
56. S. Guindon, J.F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, Syst. Biol. **59**, 307–321 (2010)
57. J.A. Nylander, MrAIC.pl. Program distributed by the author. Evolutionary Biology Centre, Uppsala University (2004)
58. F. Ronquist, J.P. Huelsenbeck, Bioinformatics **19**, 1572–1574 (2003)
59. G. Altekar, S. Dwarkadas, J.P. Huelsenbeck, F. Ronquist, Bioinformatics **20**, 407–415 (2004)
60. C. Salgado-Salazar, L.R. Jones, A. Restrepo, J.G. McEwen, Cladistics **26**, 613–624 (2010)
61. D.H. Huson, D.C. Richter, C. Rausch, T. Dezulian, M. Franz, R. Rupp, BMC Bioinf. **8**, 460 (2007)
62. R-Development-Core-Team R, *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2010)
63. M. Pagel, Nature **401**, 877–884 (1999)
64. M. Pagel, A. Meade, D. Barker, Syst. Biol. **53**, 673–684 (2004)
65. M. Pagel, Zool. Scr. **4**, 331–348 (1997)
66. L. Bellec, N. Grimsley, E. Derelle, H. Moreau, Y. Desdevises, Environ. Microbiol. Rep. **2**, 313–321 (2010)
67. A. Monier, J.-M. Claverie, H. Ogata, Genome Biol. **9**, R106 (2008)
68. D. Moreira, C. Brochier-Armanet, BMC Evol. Biol. **8**, 12 (2008)
69. G.S. Hayward, J.C. Zong, Curr. Top. Microbiol. Immunol. **312**, 1–42 (2007)
70. J.C. Zong, D.M. Ciufo, D.J. Alcendor, X. Wan, J. Nicholas, P.J. Browning, P.L. Rady, S.K. Tyring, J.M. Orenstein, C.S. Rabkin, I.J. Su, K.F. Powell, M. Croxson, K.E. Foreman, B.J. Nickoloff, S. Alkan, G.S. Hayward, J. Virol. **73**, 4156–4170 (1999)
71. K.M. Ibrahim, R.A. Nichols, G.M. Hewitt, Heredity **77**, 282–291 (1996)