

ARTICLE

Received 3 May 2013 | Accepted 2 Oct 2013 | Published 6 Nov 2013

DOI: 10.1038/ncomms3700

# Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses

Simon Roux<sup>1,2</sup>, François Enault<sup>1,2</sup>, Gisèle Bronner<sup>1,2</sup>, Daniel Vaultot<sup>3</sup>, Patrick Forterre<sup>4,5</sup> & Mart Krupovic<sup>4</sup>

Metagenomic studies have uncovered an astonishing diversity of ssDNA viruses encoding replication proteins (Reps) related to those of eukaryotic *Circoviridae*, *Geminiviridae* or *Nanoviridae*; however, exact evolutionary relationships among these viruses remain obscure. Recently, a unique chimeric virus (CHIV) genome, which has apparently emerged via recombination between ssRNA and ssDNA viruses, has been discovered. Here we report on the assembly of 13 new CHIV genomes recovered from various environments. Our results indicate a single event of capsid protein (CP) gene capture from an RNA virus in the history of this virus group. The domestication of the CP gene was followed by an unprecedented recurrent replacement of the *Rep* genes in CHIVs with distant counterparts from diverse ssDNA viruses. We suggest that parasitic and symbiotic interactions between unicellular eukaryotes were central for the emergence of CHIVs and that such turbulent evolution was primarily dictated by incongruence between the CP and Rep proteins.

<sup>1</sup>Laboratoire 'Microorganismes: Génome et Environnement', Clermont Université, Université Blaise Pascal, Clermont-Ferrand 63000, France. <sup>2</sup>CNRS UMR 6023, LMGE, Aubière 63171, France. <sup>3</sup>UPMC (Paris-06) and CNRS, UMR 7144, Station Biologique, Roscoff 29680, France. <sup>4</sup>Unité Biologie Moléculaire du Gène chez les Extrémophiles, Département de Microbiologie, Institut Pasteur, 25 rue du Dr. Roux, Paris 75015, France. <sup>5</sup>Laboratoire de Biologie Moléculaire du Gène chez les Extrémophiles, Institut de Génétique et Microbiologie, CNRS UMR 8621, Université Paris Sud, Orsay 91405, France. Correspondence and requests for materials should be addressed to M.K. (email: krupovic@pasteur.fr).

Single-stranded (ss) DNA viruses represent a rapidly expanding, diverse supergroup of economically, medically and ecologically important pathogens preying on hosts from all three domains of life. On the basis of genetic and structural properties, they are classified by the International Committee on Taxonomy of Viruses into eight families—*Anelloviridae*, *Bidnaviridae*, *Circoviridae*, *Geminiviridae*, *Inoviridae*, *Microviridae*, *Nanoviridae* and *Parvoviridae*<sup>1</sup>—whereas some groups still await proper taxonomical assessment<sup>2–4</sup>. ssDNA viruses infecting plants (geminiviruses and nanoviruses) and animals (anelloviruses, circoviruses and parvoviruses) were in the spotlight of extensive research for many years due to their direct effect on the well-being of humans. Recently, a previously unsuspected facet of the ssDNA viruses as important factors in the global ecosystems has come to light; viruses with ssDNA genomes have been repeatedly isolated from diverse environments, including extreme geothermal<sup>2</sup> and hypersaline habitats<sup>3,5</sup>, soil<sup>6</sup>, freshwater and marine ecosystems<sup>7–9</sup>.

Whereas some bacterial and archaeal ssDNA viruses display filamentous or pleomorphic (variable appearance) virion morphologies<sup>2,3,10</sup>, all eukaryotic ones (namely *Anelloviridae*, *Bidnaviridae*, *Circoviridae*, *Geminiviridae*, *Nanoviridae* and *Parvoviridae*) pack their genomes into small icosahedral capsids, constructed from multiple copies of a single (for example, geminiviruses) or several (for example, parvoviruses) nearly identical capsid proteins (CP)<sup>11</sup>. In all cases when high-resolution structural information is available, the CPs of ssDNA viruses were found to display a jelly-roll (antiparallel eight-stranded  $\beta$ -barrel) fold, which is also found in the vast majority of icosahedral positive-sense ssRNA viruses infecting eukaryotic hosts<sup>12,13</sup>. At the sequence level, however, the similarity between the CPs of viruses belonging to different families is not recognizable. Another feature that is common to eukaryotic ssDNA viruses is the mechanism of genome replication; the vast majority of these viruses are believed to replicate their genomes via the rolling-circle (RC) or catalytically similar rolling-hairpin mechanism mediated by homologous virus-encoded RC replication initiation proteins (RC-Rep)<sup>13,14</sup>. In this respect, ssDNA viruses resemble prokaryotic RC plasmids, pointing towards a possible evolutionary link between these two types of mobile genetic elements<sup>13–15</sup>. A characteristic feature of RC-Reps of eukaryotic ssDNA viruses is the presence of the superfamily 3 helicase (S3H) domain<sup>16,17</sup>, which is fused carboxy-terminally to the catalytic nuclease domain encompassing three signature motifs found in all prokaryotic and eukaryotic virus and plasmid RC-Reps<sup>14</sup>. As opposed to CPs, RC-Reps of eukaryotic viruses display actual sequence similarity and RC-Rep-based phylogenies recapitulate the major taxonomic groups defined by the International Committee on Taxonomy of Viruses<sup>1,18,19</sup>. However, it should be noted that not all eukaryotic ssDNA viruses possess genes for canonical RC-Reps; for example, anelloviruses—even though believed to replicate via RC mechanism—do not encode a protein that would contain the entire set of motifs characteristic to RC-Reps<sup>17</sup>.

The origin(s) and evolutionary relationships between ssDNA viruses belonging to different families remain obscure. Structural similarity between the CPs of bacterial microviruses and eukaryotic parvoviruses, circoviruses and geminiviruses<sup>11</sup> was suggested to testify for the common origin of these viruses<sup>20</sup>. Alternatively, similarity between the RC-Reps of ssDNA viruses and prokaryotic plasmids on the one hand<sup>14,15,21,22</sup> and structural similarity between the CPs of viruses with ssDNA and ssRNA genomes on the other<sup>13</sup> led to the proposal that different groups of ssDNA viruses have emerged from plasmids by acquisition of CP-coding genes from RNA viruses, possibly on multiple independent occasions<sup>12,13,23</sup>. Indeed, both homologous

and illegitimate recombination have important roles in driving the evolution of ssDNA viruses<sup>19</sup>.

During the past few years, numerous studies on uncultivated viral communities using metagenomic approaches have revealed that genetic diversity of ssDNA viruses is much greater than originally recognized (reviewed in refs 17,18). Many of these uncultivated viruses are related to members of the bacteriophage family *Microviridae*<sup>9</sup>, but perhaps even larger number encode RC-Reps displaying phylogenetic affinity to one of three families of eukaryotic ssDNA viruses—*Circoviridae*, *Geminiviridae* and *Nanoviridae* (for example, see refs 24–27). Interestingly, instead of encoding genes for corresponding CPs (circo-, gemini- and nano-like), these viruses typically bear open-reading frames that do not share appreciable similarity with sequences in the databases. Potentially, the lack of recognizable sequence similarity might be caused by the extremely high mutation rates characteristic to ssDNA viruses<sup>28,29</sup>. Thus, to navigate in the constantly increasing pool of environmental viral genomes, RC-Reps are often used as markers for classification of the uncultivated ssDNA viruses.

Recently, Diemer and Stedman<sup>30</sup> have described a novel chimeric viral (CHIV) ssDNA genome recovered from a hot, acidic Boiling Springs Lake (BSL), USA. Whereas the RC-Rep of the virus was most similar to those of circoviruses, the CP was highly similar to the CPs of ssRNA viruses of the family *Tombusviridae* and two unclassified oomycete-infecting viruses, *Sclerophthora macrospora* virus A (SmV-A) and *Plasmopara halstedii* virus A (PhV-A)<sup>30</sup>. Notably, the tombusvirus-like CP topology has not been previously observed for any DNA virus, suggesting that the virus has emerged via recombination between a DNA and an RNA virus<sup>31</sup>. The validity of the assembled viral genome, tentatively named the RNA–DNA hybrid virus (BSL\_RDHV), and its presence in the lake sediment pore water were confirmed by PCR amplification<sup>30</sup>. Importantly, the finding that RNA and DNA viruses recombine to produce novel chimeric entities rationalizes some of the puzzles of the virosphere and allows assessing new hypotheses on the origin and evolution of different viral groups<sup>12,13</sup>.

Here we report on the assembly of 13 new CHIV genomes recovered from various environments, and encoding tombusvirus-like CPs and, unexpectedly, diverse RC-Reps related to the corresponding proteins of eukaryotic ssDNA viruses belonging to three different families. We show that the history of this virus group involved a unique event of CP gene capture from an RNA virus, followed by an unprecedented recurrent replacement of the Rep genes in CHIVs with distant counterparts from diverse ssDNA viruses. Frequent exchange of Rep genes described here blurs the borders between the major groups of eukaryotic ssDNA viruses and suggests that Reps represent an inadequate marker for tracing their evolutionary history. Finally, we suggest that parasitic and symbiotic interactions between unicellular eukaryotes were central for the emergence of CHIVs.

## Results

**New CHIVs.** To get further insight into the diversity and evolution of CHIVs, we have assembled sequence reads from 103 publicly available viromes and searched the resultant contigs for co-occurrence of genes encoding RC-Reps and RNA virus-like CPs (Supplementary Data 1). As a result, nine contigs were assembled from viromes derived from atmospheric<sup>32</sup> and aquatic<sup>26,33,34</sup> samples. As ssDNA viruses are known to integrate into the genomes of their hosts<sup>35,36</sup>, we also searched for the presence of CHIVs in the eukaryotic genome databases. The latter approach yielded four additional contigs matching our criteria. Three of these represented contigs from two different

whole-genome shotgun (WGS) libraries of marine photosynthetic picoeukaryote populations dominated by the green alga *Bathycoccus*<sup>37</sup>, whereas the fourth one was from a WGS library of *Astrammmina rara*, a foraminiferan protist<sup>38</sup>.

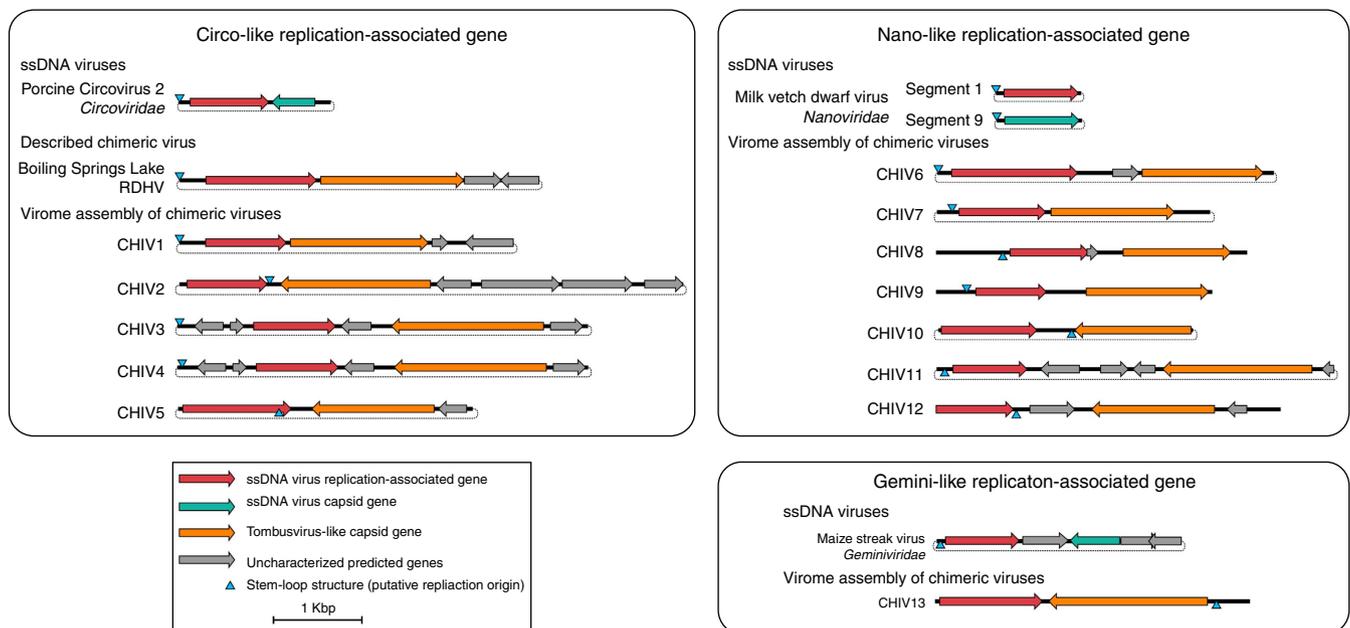
General characteristics of the 13 CHIV genomes (CHIV1–13) obtained by the two approaches are summarized in Fig. 1 and Supplementary Data 2. In accordance with the experimentally verified topology of the BSL\_RDHV genome<sup>30</sup>, most (9 out of 13) of the CHIV contigs obtained here are circular. Importantly, the potential stem loops containing nonanucleotide sequences, which serve as origins of replication in ssDNA viruses with circular genomes<sup>17</sup>, are readily identifiable in proximity of the *RC-Rep* genes in all CHIV genomes (Supplementary Data 2 and Fig. 1). Besides the *CP* and *RC-Rep* genes, some of the CHIVs are predicted to contain up to four additional open-reading frames. However, sequence analysis does not offer any insight into their possible functions.

**Emergence of CHIVs is a rare event.** All CHIVs encode putative CPs related to those of tombusviruses, to the exclusion of all other groups of RNA viruses. We note that recent exploration of the ssDNA virus diversity associated with dragonflies revealed a viral genome, DfCyclV, encoding a putative protein with weak but significant similarity to the CP of satellite tobacco necrosis virus<sup>25</sup>. The authors concluded that DfCyclV might be a CHIV with a circovirus-like *RC-Rep* and a tombusvirus-like CP. However, the satellite tobacco necrosis virus CP is radically different in sequence and structure from those of tombusviruses and most closely resembles the CPs of geminiviruses<sup>21</sup>. Thus, in our opinion, DfCyclV should not be confused with BSL\_RDHV-like CHIVs.

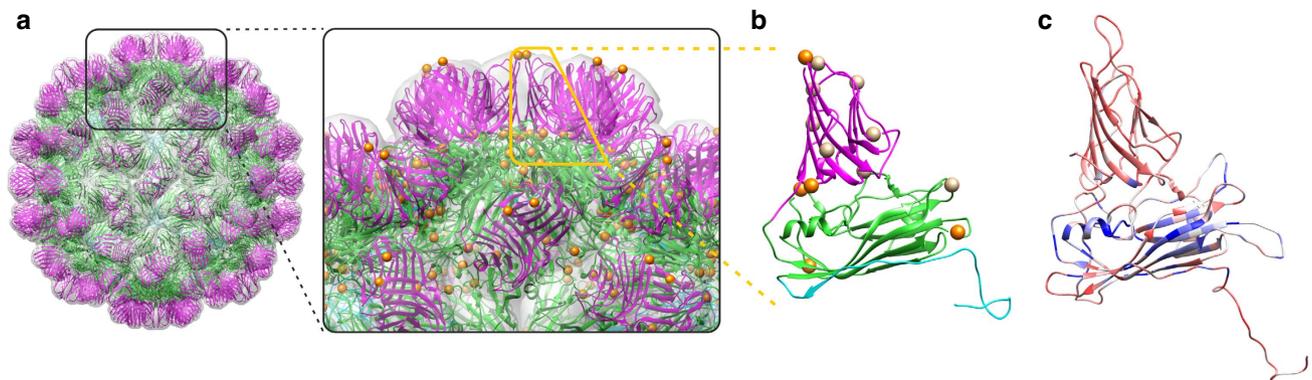
Members of the family *Tombusviridae* have positive-sense ssRNA genomes and infect a variety of land plants<sup>39</sup>, although several tombusviruses have also been isolated from freshwater samples<sup>40</sup>. Viruses belonging to *Tombusviridae* genera *Aureusvirus*, *Avenavirus*, *Carmovirus*, *Dianthovirus* and *Tombusvirus* possess icosahedral virions with a granular surface.

The latter property is determined by a unique domain organization of the CPs of these viruses. Each CP subunit consists of three distinct domains: the amino-terminal RNA-binding (R) domain facing the interior of the virion, the shell (S) domain central for the assembly of the icosahedral capsid and the C-terminal projection (P) domain, which faces away from the capsid surface, giving the virion its granular appearance (Fig. 2a,b). Outside of the *Tombusviridae*, the same CP domain organization is expected (based on sequence similarity) only for two recently isolated unclassified oomycete-infecting ssRNA viruses, SmV-A and PhV-A<sup>41,42</sup>.

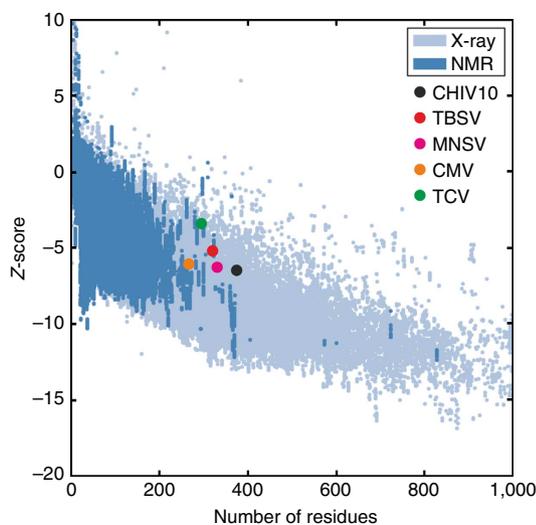
To better understand the relationship between the CPs of ssRNA viruses and CHIVs, we built a three-dimension model of a representative CHIV CP; CHIV10 (Airborne\_IC2) was chosen for this purpose (Fig. 2c). In accordance with previous predictions<sup>30</sup>, good stereochemical quality of the obtained structural model (Fig. 3) confirms that the CPs of CHIVs are likely to display the same structural fold and domain organization as those of tombusviral CPs. Comparison of the 14 CHIV CPs (13 new and 1 from BSL\_RDHV) in the context of their tertiary structures reveals that the most conserved part of these proteins corresponds to the S-domain, whereas the R- and P-domains are much more variable (Fig. 2c, Supplementary Fig. S1). Similar pattern of conservation has been also observed in tombusviruses<sup>43</sup>. Closer examination of the multiple alignment of CHIV, tombusviral and oomycete-infecting virus (SmV-A and PhV-A) CP sequences shows that CHIV CPs are more closely related to the proteins of SmV-A and PhV-A than they are to the CPs of tombusviruses. Five unique insertions, not present in tombusviral CPs, are shared between the CPs of CHIVs, SmV-A/PhV-A and the related sequences from the Lake Needwood RNA virome (indicated with orange spheres in Fig. 2a,b; see also Supplementary Fig. S1), which we consider as synapomorphies testifying for the common evolutionary history of these proteins. Furthermore, unlike in tombusviruses, but similar to that in SmV-A/PhV-A, CHIV capsids are not likely to be stabilized by calcium ions; none of the CHIV CPs contains the calcium-binding motifs, as has been also noted for BSL\_RDHV<sup>30</sup>. Finally, eight species-specific insertions are present



**Figure 1 | Genomic maps of CHIVs and representative reference genomes.** CHIVs are grouped according to the type of Rep they encode. Arrows denote open-reading frames. The colour key is provided in the figure. The location and orientation of the potential stem loops containing nonanucleotide motifs are indicated with light blue triangles. CHIV13 genome is reverse-complemented for more convenient representation.



**Figure 2 | Insights into the structure of CHIVs.** (a) Structure of the MNSV (PDB ID:2ZAH) is shown to illustrate the contribution of the distinct CP domains to the virion organization and the position of these domains in the capsid surface lattice. The P-domain, magenta; S-domain, green; R-domain, cyan. On the right is a zoom-in on one of the capsid areas, where locations of the conserved insertions present in the CPs of CHIVs, as well as viruses SmV-A and PhV-A are indicated with orange spheres. (b) Structure of the MNSV CP. The P-, S- and R-domains, as well as the locations of conserved insertions are coloured as in a. In addition, the locations of species-specific sequence insertions present in the CPs of only some CHIVs are indicated with grey spheres. (c) Structural model of the CHIV10 CP. The colouring represents sequence conservation among CHIV CPs (red, least conserved; blue, most conserved).



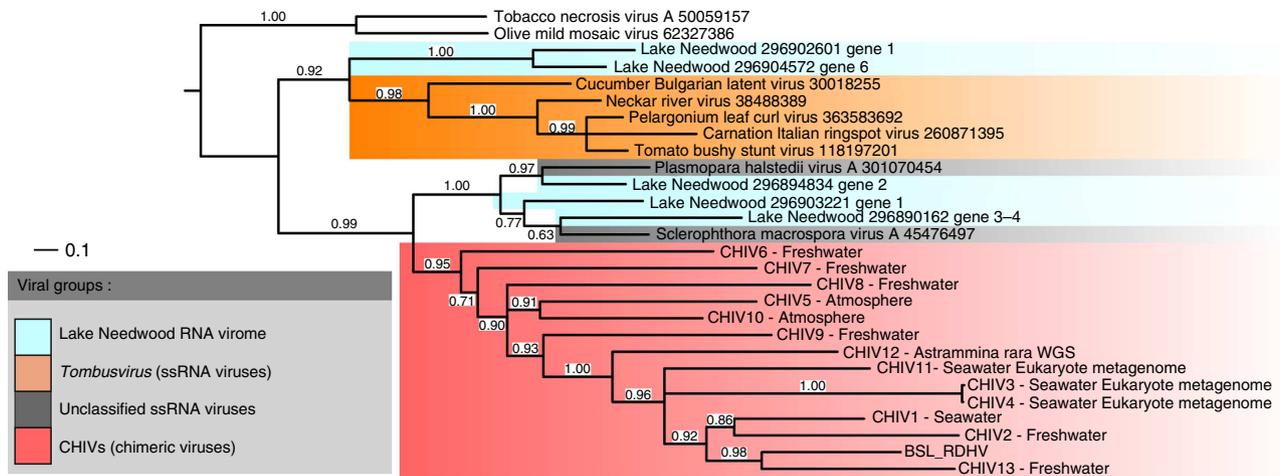
**Figure 3 | Quality assessment of the three-dimensional model of the CHIV10 CP.** Quality of the generated model along with that of structural homologues used for modelling (see Methods section) was evaluated using PsoSA-web at <https://prosa.services.came.sbg.ac.at/prosa.php>. The calculated quality (Z) scores (closed circles) are displayed in the context of the Z-scores of all experimentally determined protein structures available in the Protein Data Bank. Every dot represents a distinct structure solved by X-ray crystallography (light blue) or NMR (dark blue). TBSV, tomato bushy stunt virus; MNSV, melon necrotic spot virus; CMV, carnation mottle virus; TCV, turnip crinkle virus.

in the CPs of certain CHIVs (grey spheres in Fig. 2b, see also Supplementary Fig. S1). Most of them are located within the P-domains. Importantly, alterations within the P-domain are less likely to interfere with capsid formation, which is primarily orchestrated by interactions within the S-domain. We hypothesize that the P-domain is involved in virus–host interaction (possibly host recognition), which would explain its greater variability promoted by a constant arms race between the virus and the host<sup>44</sup>.

To learn on how many independent occasions tombusvirus-like CP genes were captured by DNA viruses, we performed a maximum-likelihood phylogenetic analysis of the CHIV, tombusviral and SmV-A/PhV-A CP proteins (Fig. 4). In addition,

the data set was supplemented with tombusvirus-like CP sequences recovered from the RNA virome obtained from Lake Needwood<sup>45</sup>. Notably, in none of the data sets containing information about both RNA and DNA virus communities present in the same environmental sample<sup>34</sup> could we detect both CHIVs and tombus-like RNA viruses (in the DNA and RNA fractions, respectively), pointing towards their divergent distribution. The tombusvirus sequences form a well-supported monophyletic clade. Interestingly, all CHIVs cluster together as a sister group to the CPs of SmV-A/PhV-A (Fig. 4). Monophyly of the CHIV CPs and the fact that no other RNA virus-like CPs were found in association with RC-Reps suggest that transfer of a CP gene between RNA and DNA viruses was a unique event and that emergence of CHIVs is likely to be rare.

**Polyphyly of RC-Reps in CHIVs.** Sequence analysis of CHIV RC-Reps reveals a domain organization typical of eukaryotic ssDNA viruses, with the N-terminal nuclease domain and the C-terminal S3H domain<sup>14,17</sup>. The three signature motifs of the nuclease domain are readily identifiable in all CHIV RC-Reps, whereas the S3H motifs are conserved in all but two proteins—Walker B motif could not be mapped in the RC-Reps of CHIV6 and CHIV12 (Table 1). Previous analysis of the RC-Rep encoded by BSL\_RDHV showed that it is most closely related to those of circoviruses<sup>30</sup>. Unexpectedly, BLASTp analysis performed in this study reveals differential affinity of the CHIV RC-Reps to the corresponding proteins from three major groups of eukaryotic ssDNA viruses. The latter observation is confirmed by phylogenetic analysis of RC-Reps encoded by CHIVs, circoviruses, nanoviruses and geminiviruses (Fig. 5). Similar to that in the case of BSL\_RDHV, five CHIVs (CHIV1–5) cluster with circoviruses. CHIV6–12 form a well-supported phylogenetic clade with nanoviruses, whereas CHIV13 branches together with geminiviruses, separately from the rest of CHIVs (Fig. 5). The significance of a tree topology can be assessed using a constrained tree approach, as demonstrated previously for other viruses<sup>46</sup>. To verify the validity of the obtained grouping of CHIV RC-Reps, the likelihood of the original tree (Fig. 5) was compared with the likelihood of a tree constrained for CHIV monophyly (see Methods section). In this analysis, the monophyly is unequivocally rejected (Table 2) at a statistically significant level ( $P$ -value  $< 0.001$ ), confirming the polyphyly of CHIV RC-Reps. By contrast, the



**Figure 4 | Phylogenetic tree of tombusvirus-like CPs.** CHIVs are highlighted in red, tombusviruses in orange and unclassified ssRNA viruses are either in grey when isolated, or in blue when assembled from Lake Needwood RNA virome. Tobacco necrosis virus A and Olive mild mosaic virus, both members of the genus *Necrovirus* within *Tombusviridae*, have CPs lacking the P-domain and were used as external group. Numbers at the branch points represent SH-like local support values (based on 1,000 resamples), and nodes with scores <0.5 were collapsed. NCBI GI numbers are indicated for all reference sequences.

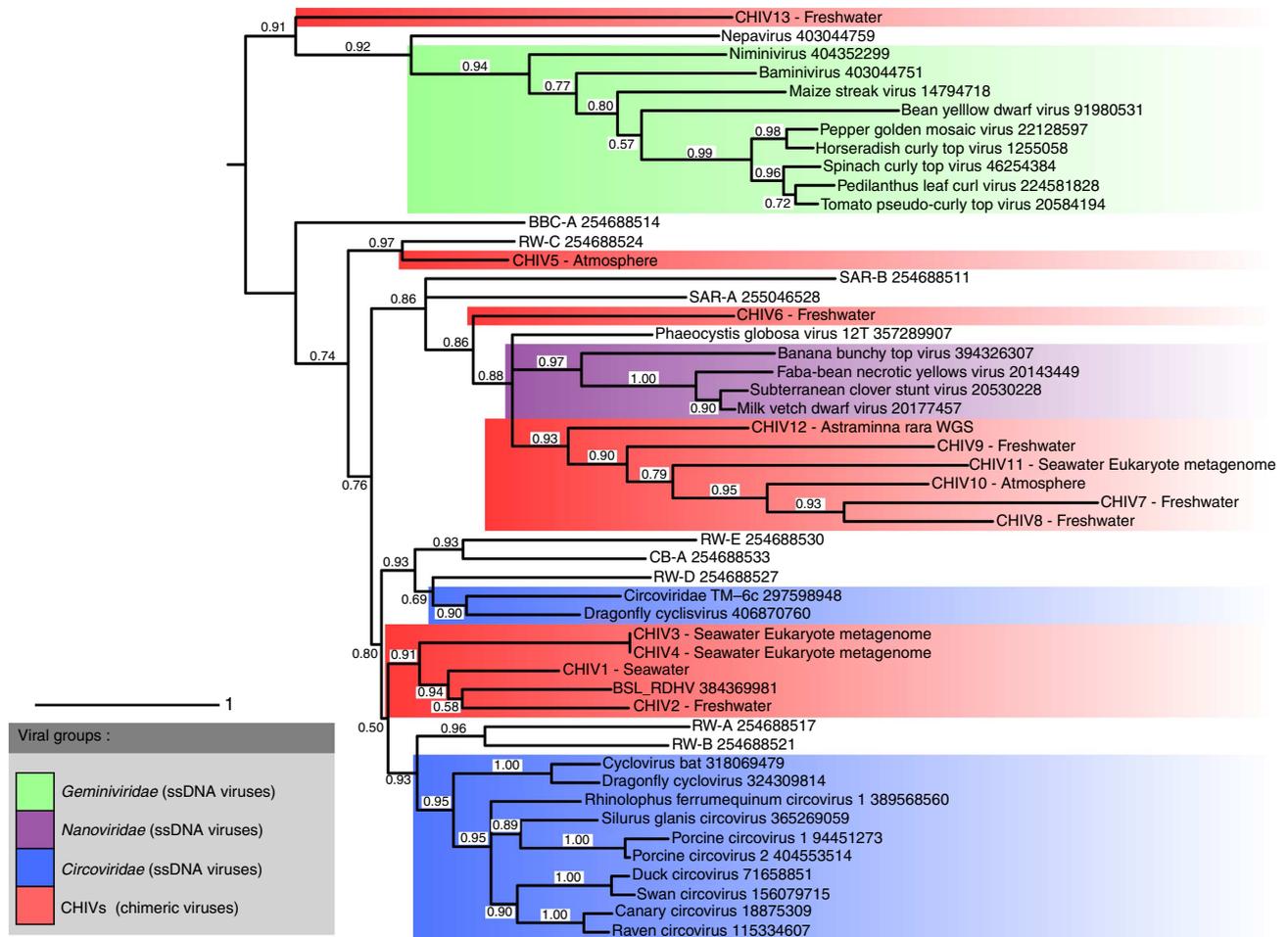
**Table 1 | RCR and S3H motifs of CHIV RC-Reps.**

	RCR motifs			S3H motifs		
	I	II	III	Walker-A	Walker-B	Walker-C
Reference viruses*						
<i>Circoviridae</i> and cycloviruses	[VC]FT[LIW]NN	[Px]HLQG	YC[Sx]JK	G[Px][PST][GC]xGKS	[VI][IUML]DDF	UTS[NE]
<i>Nanoviridae</i>	[VCx]FT[LI]N[FYN]	xHUQG	Y[CAS]xK	G[PS]xG[GN]EGK[TS]	[VIW][UAC][FIM]D[IVF]	V[FMII][AC]N
<i>Geminiviridae</i>	FLTYP[S]x	[Px]H[Lx]H[VAC]	Y[UAC]xK	Gx[ST]R[TI]GK[TS]	[VI][IV]DD[VI]	UL[Cx]N
CHIVs						
BSL RDHV	CITVNN	KHLQV	YCTK	GATGTGKS	IIDDY	ITCP
CHIV1	CFTLNN	RHLQG	YCSK	GPTGTGKT	IIDDY	ITAP
CHIV2	VLVLNN	IHLQG	YITK	GETGQGS	IIDDY	ITTP
CHIV3	CITHNN	KHIQA	YCIK	GETGTGKS	IIDDV	FTAP
CHIV4	CITHNN	KHIQA	YCIK	GETGTGKS	IIDDV	FTAP
CHIV5	VFTLNN	PHIQG	YCSK	GDTGTGKS	VINDF	ITSS
CHIV6	CFTYYY	PHLQS	YCKK	PKGGNGKT	—	VFSN
CHIV7	DFRLNQ	IHFQG	YASK	VKGCQKQ	IFLDL	VFTN
CHIV8	MFDWRH	DHYQG	YASK	PVGNGKT	TLIDL	VFTN
CHIV9	LITYKS	PHTHA	YLSK	SKGNAGKS	IIFDI	VMSN
CHIV10	CCVYDF	KHFQG	YTMK	EKNIGKT	YLIDM	IFTN
CHIV11	SWTWNK	LHYQG	YCMK	DTGCTGKS	YIFDI	VFSN
CHIV12	CFTFNN	PHLQG	YCSK	EVGQLGKT	—	VLSN
CHIV13	HLTYKT	DHTHF	YHKK	GSTNTGKT	VFDDM	FTSN

\* - conserved RCR and S3H motifs of reference virus groups were retrieved from ref. 17. BSL, Boiling Springs Lake; CHIV, chimeric virus; RC-Reps, RC replication initiation protein; S3H, superfamily 3 helicase; U, bulky hydrophobic residues (I, L, V, M, F, Y, W); x, any amino acid residue.

constrained tree strongly enforces the monophyly of CHIV CPs and cannot be rejected by statistical tests (Table 2). Such phylogenetic distribution of CHIV RC-Reps is in stark contrast with the monophyly of the CHIV CPs. Indeed, the CHIV pairs that are close on the CP tree fall into different clades on the RC-Rep phylogeny. For example, the three CHIVs recovered from the WGS library of the photosynthetic picoeukaryotes fall into two different groups (CHIV3 and CHIV4 encode circovirus-like RC-Reps, whereas CHIV11 has a nanovirus-like protein), despite the fact that their CPs cluster together (Fig. 4). Similarly, the CP of CHIV13 is closely related to the corresponding BSL\_RDHV protein, but their RC-Reps group with geminiviruses and circoviruses, respectively (Fig. 5).

To compare the evolutionary patterns of CHIVs, circoviruses, nanoviruses, geminiviruses and tombusviruses, we have plotted the pairwise distances calculated for CPs from the representative members within each taxon against the corresponding distances between their replication proteins (Reps; Fig. 6a). We found that in circoviruses, nanoviruses, geminiviruses and tombusviruses, the Repls are considerably less divergent than the corresponding CPs. Strikingly, the pattern is the opposite in CHIVs; RC-Reps are much more divergent than in any other virus taxon. In combination with the results of phylogenetic analysis (Fig. 5), such sequence divergence of CHIV RC-Reps is most consistent with multiple independent events of RC-Rep gene replacement in different CHIVs.



**Figure 5 | Phylogenetic analysis of the CHIV RC-Reps.** CHIVs are highlighted in red, circoviruses in blue, nanoviruses in purple and geminiviruses in green. Environmental sequences amplified from Reclaimed Water (RW), Chesapeake Bay (CB) or British Columbia (BBC) were taken as additional references<sup>24,34</sup>, as well as RC-Rep gene from double-stranded DNA (dsDNA) algal *Phaeocystis globosa* virus 12T. Numbers at the branch points represent SH-like local support values (based on 1,000 resamples), and nodes with scores <0.5 were collapsed. NCBI GI numbers are indicated for all reference sequences.

**Table 2 | Statistical analysis of constrained trees.**

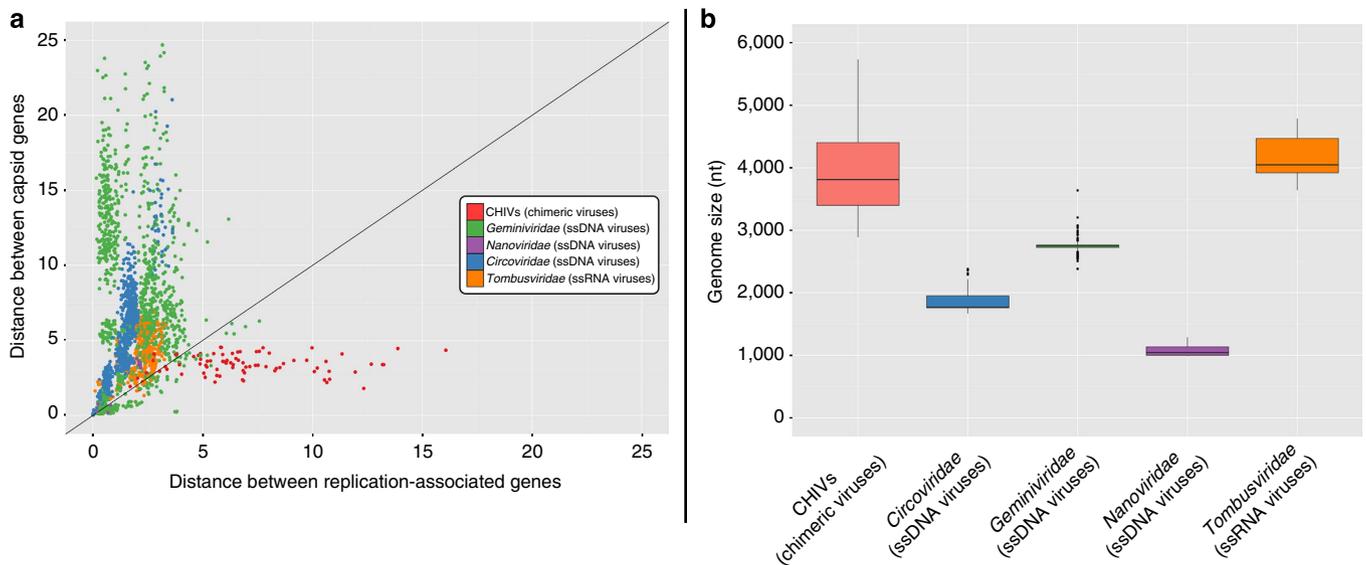
	ELW	AU test	Likelihood	Assessment
<i>Coat protein</i>				
Unconstrained tree	1	0.254	-24,920.74	—
Constrained tree (CHIV clade)	1	0.235	-24,920.74	—
<i>RC-Rep protein</i>				
Unconstrained tree	1	0.274	-27,032.60	—
Constrained tree (CHIV clade)	0	0	-27,170.82	Rejected

AU, approximately unbiased; CHIV, chimeric virus; ELW, expected-likelihood weight; RC-Reps, RC replication initiation protein.

**Unicellular algae as recombination hotspots.** Although viromes studied here were assembled from a wide range of biomes (Supplementary Data 1), CHIVs are exclusively retrieved from aquatic and atmospheric environments. Similarly, when microbial metagenomes are considered, CHIVs once again are identified only in aquatic samples. Three CHIV genomes (two of which are very similar, CHIV3 and CHIV4) are detected in two different samples enriched for the photosynthetic unicellular alga

*Bathycoccus*, pointing towards potential association between algae and CHIVs. The fourth CHIV genome associated with aquatic microbes is found in the WGS library of *A. rara*. It is worth noting that foraminiferans are often engaged in endosymbiotic relationships and were found to host unicellular algae belonging to diverse lineages, including green algae, red algae, diatoms and dinoflagellates<sup>47</sup>. Consequently, it is possible that the CHIV contig associated with *A. rara* derives from an algal symbiont, rather than *A. rara* itself. At any rate, the association of different CHIV contigs with two different types of eukaryotes raises an intriguing possibility that unicellular eukaryotes serve as hosts for at least some CHIVs.

Interestingly, we identified a close homologue (AET73220;  $E = 4e - 29$ , 35% identity) of CHIV12 RC-Rep (but not the CP) encoded in the genome of a giant double-stranded DNA virus, PgV-12T, infecting *Phaeocystis globosa*<sup>48</sup>, a photosynthetic unicellular alga. It has been recently demonstrated that satellite viruses and transposons integrate into the genome of the Lentille virus, a relative of mimiviruses<sup>49</sup>. It is tempting to speculate that ssDNA viruses and derived elements might represent a new class of molecular parasites preying on giant double-stranded DNA viruses. Regardless, the presence of the RC-Rep gene in the genome of PgV-12T lends additional support to the hypothesis that unicellular algae may host at least some of the CHIVs. More generally, parasitic and symbiotic relationships involving



**Figure 6 | Comparison of CHIVs with other ssDNA viruses and tombusviruses.** (a) Evolutionary distance (Jones-Taylor-Thornton (JTT) model) for RC-Rep and CP sequences assessed between pairs of genomes within ssDNA and ssRNA families, and between CHIVs. (b) Box plot of genome size distribution in CHIVs (14 genomes), circoviruses (67), geminiviruses (464), nanoviruses (4) and tombusviruses (48). Whiskers correspond to the R ggplot library geom\_boxplot function default paramers (for upper whisker: the highest value that is within 1.5\*Inter-Quartile Range of the hinge, and for lower whisker: the smallest value that is within 1.5\*Inter-Quartile Range of the hinge). Any value outside these whiskers is considered as an outlier and displayed as a dot.

unicellular algae are highly prevalent in aquatic environments<sup>50</sup> and might be central for the emergence of new virus types, such as CHIVs, by providing a unique environment accessible for viruses infecting phylogenetically distant hosts. Such co-localization of various genetic elements of distinct origins and histories could also explain the evolutionary relationships between RC-Reps of prokaryotic plasmids and eukaryotic ssDNA viruses<sup>12,13,15,21,22</sup>.

## Discussion

Recombination is known to have an important role in the evolution of eukaryotic ssDNA viruses<sup>13,19</sup>. However, interfamilial gene exchange has not been convincingly demonstrated for these viruses, suggesting that such recombination might be either uncommon or the recombinants are rarely retained in the population. In this light, pervasive exchange of *RC-Rep* genes in CHIVs is surprising. We hypothesize that the unusually frequent *RC-Rep* gene transfer in the CHIV lineage could have been instigated by incongruences between the capsid and *RC-Rep* proteins in the ancestral CHIV. It appears reasonable to assume that CP and *RC-Rep*, which evolved in the contexts of RNA and DNA viral genomes, respectively, would not immediately form a perfect match. Thus, *RC-Rep* genes could have been exchanged as long as the CP-Rep combination is not optimal. However, once the CP and the *RC-Rep* genes are sufficiently adapted to each other (that is, further ‘sampling’ decreases fitness) and/or viruses occupy a specific niche where ‘sampling’ is no longer possible, such high rate of gene exchange is expected to transit to a more conservative mode observed in other eukaryotic ssDNA viruses.

Metagenomic studies have recently uncovered the unsuspected diversity of ssDNA viruses, many of which encode RC-Reps similar to those of geminiviruses, nanoviruses and, perhaps most commonly, circoviruses<sup>17,18</sup>. However, their CP genes are typically beyond recognition using sequence-based approaches, opening a possibility that these uncultured viruses represent highly divergent yet genuine members of the corresponding viral

families. By contrast, CHIVs described here—despite being scattered throughout the RC-Rep phylogeny (Fig. 5)—all share a CP gene, which they apparently inherited from a common ancestor (Fig. 4). Importantly, tombusvirus-like CP gene is not the only feature that distinguishes CHIVs from the three families of eukaryotic viruses mentioned above. CHIV genomes are also significantly larger than those of geminiviruses, nanoviruses and circoviruses, and are close in size to the ssRNA genomes of tombusviruses (Fig. 6b). Consequently, capsids larger than those of ssDNA viruses would be required to package such genomes. Interestingly, mechanical properties, such as persistence length, of ssRNA and ssDNA molecules are similar<sup>51</sup>, indicating that tombusvirus-like capsids would be well fitted to accommodate the larger genomes of CHIVs.

Where do viruses with RNA virus-like capsids, DNA genomes and RC-Rep diversity spanning the major groups of eukaryotic ssDNA viruses fit in the virosphere? Obviously, CHIVs cannot be neatly placed into any one of the established groups of ssDNA viruses. Furthermore, evidence that *RC-Rep* genes can be exchanged between unrelated viruses blurs the borders between the major groups of eukaryotic ssDNA viruses and renders the RC-Rep-based classification of the uncultured ssDNA viruses into the circo-, nano- or gemini-like groups obsolete. Indeed, CHIVs with circovirus-like RC-Reps are as similar to circoviruses (that is, circovirus-like)<sup>30</sup>, as they are to tombusviruses<sup>12</sup>. Recognizing the limits of the RC-Rep-based approach in classifying uncultured ssDNA viruses, Rosario *et al.*<sup>17</sup> have recently proposed an alternative classification scheme based on a combination of various genomic properties of these viruses. According to the new scheme, viruses are categorized into eight groups (I–VIII) based on their genome orientation, the location of the intergenic region containing the potential stem loop structure, as well as the orientation of the nonanucleotide motif with respect to the *RC-Rep* gene<sup>17</sup>. The diversity of genome organizations observed in CHIVs spans six of the eight proposed groups (Fig. 1 and Supplementary Data 2), suggesting that such classification scheme might not prove to be practical.

More generally, none of the viral genes taken separately can adequately represent viral history<sup>52</sup>, especially so in the light of rampant horizontal gene exchange in the viral world<sup>53</sup>. Genetic mosaicism has been previously pointed out as a factor impeding meaningful classification of tailed bacteriophages (order Caudovirales)<sup>54</sup>. However, the coding capacity of tailed bacteriophages is typically large enough to accommodate a representative core gene set<sup>55,56</sup> sufficient for hierarchical clustering of these viruses into biologically significant subdivisions<sup>57,58</sup>. For viruses with small genomes, on the other hand, the effect of horizontal gene transfer on the 'identity' of a viral group is considerably more acute. Thus, eukaryotic ssDNA viruses, which usually encode only a handful of proteins, in our opinion, represent a clear-cut case of organisms for which ancient evolutionary history cannot be reconstructed employing whole-genome approaches.

In a situation where objective means of virus classification are not applicable, a different—even if suboptimal—solution has to be sought. One way would be to classify CHIVs (and ssDNA viruses in general) based on their Reps into different viral groups, neglecting the history and nature of their CP genes. Such approach would be coherent with the Baltimore classification (that is, all viruses with ssDNA genomes would be collected together). However, such grouping would be inconsistent with our finding that RC-Reps were replaced on multiple occasions within the CHIV group. Furthermore, such scheme would be blind to the inferred structural uniformity of this viral group: all CHIVs are likely to possess similar capsids, considerably larger than those of ssDNA viruses but related in size and appearance to the capsids of tombus-like viruses (Fig. 2). Notably, CPs are hallmarks of viruses and are less likely to leave the realm of virosphere than Reps that are often exchanged between unrelated viruses, plasmids and cellular chromosomes<sup>59</sup>. Thus, alternative approach would involve virus classification based on CPs. Which of these two classification schemes will prove to be more practical remains to be seen. Difficulties with classification of new ssDNA virus groups notwithstanding exploration of the viral world has presented valuable insights into the origin and evolution of viruses. It is now obvious that the virosphere is only gradually revealing its secrets—the more we sample the virosphere, the more unexpected connections we uncover between viruses that once were considered unrelated.

**Note added in proof:** Following the revision of this article, Hewson *et al.* have described the identification of a new CHIV genome in samples collected from Oneida and Cayuga lakes (upstate New York, USA)<sup>71</sup>, further expanding our knowledge on the genetic diversity and environmental distribution of this peculiar group of chimeric viruses. Interestingly, the new ssDNA virus appears to be associated with planktonic crustaceans of the genus *Daphnia*. Genomic analysis showed that the new CHIV genome encodes a circovirus-like RC-Rep and displays an ambisense genome orientation, like in the case of CHIV3 and CHIV4, which are associated with seawater picoplankton.

## Methods

**Detection of CHIVs in assembled viromes.** A set of 103 published viromes available in public databases were downloaded and used in this study. These viromes were obtained from viral communities associated with different types of aquatic samples (freshwater, seawater and hypersaline ponds), eukaryote-associated flora (the human gut, saliva, lung, coral and fish), as well as with more peculiar biomes like microbialites or atmospheric samples (Supplementary Data 1). All viromes were assembled with Newbler 2.6 (454 Life Sciences), with the following parameters: 98% similarity over 35 bp. A BLASTx search was computed to detect contigs containing genes similar to those of RNA viruses (extracted from the NCBI protein database on Aug 2012). Genes were predicted with MetaGeneAnnotator<sup>60</sup> for all contigs that were found to encode putative RNA virus capsid-like proteins (threshold of 50 on bitscore and 0.001 on e-value). Contigs containing

at least two genes, one similar to an RNA virus capsid gene and one to the RC-Rep gene were considered as CHIVs (Supplementary Data 1). All of these contigs presented coverage  $>7\times$ , and up to  $395\times$  (Supplementary Data 2).

**Screening of WGS libraries.** Different databases from the NCBI were screened for the presence of CHIVs based on the ten CPs from CHIVs (the nine contigs assembled in this study and the BSL\_RDHV genome<sup>30</sup>). Searches against genomic survey sequence, WGS and high-throughput genomic sequence libraries were performed using tBLASTn, whereas BLASTp was used to compare CHIV CP sequences to metagenomic proteins (env\_nr). Putative CHIVs were detected in metagenomes targeting the small eukaryotic fraction in coastal upwelling waters off central Chile (NCBI GI:372349332 and 393314887)<sup>37</sup>. Reads from these two data sets were assembled with the same pipeline as the viromes, and three putative CHIV genomes were obtained. In addition, putative CHIV genome was retrieved from a WGS project of a foraminifera, *Astrammmina rara* (NCBI Bioproject PRJNA47149; Contig ADN101003178)<sup>38</sup>.

**Structural modelling and model quality assessment.** The three-dimensional model of the putative CP of CHIV10 was constructed using the advanced multi-template approach with MODELLER v9.9 (ref. 61). X-ray structures of tomato bushy stunt virus (TBSV; Protein Data Bank (PDB) ID: 2TBV), melon necrotic spot virus (MNSV; PDB ID: 2ZAH), carnation mottle virus (CMV; PDB ID: 1OPO) and turnip crinkle virus (TCV; PDB ID: 3ZXA) were used as templates. Sequence of CHIV10 CP was aligned with the corresponding sequences of TBSV, MNSV, CMV and TCV, and the resultant alignment was used to build a three-dimensional model of the putative CP of CHIV10. The initial model was optimized via multiple rounds of loop refinement with MODELLER. The stereochemical quality of the model was then assessed with ProSA-web<sup>62</sup>. ProSA-web quality (Z) score for the CHIV10 model was calculated to be  $-6.49$ , which is similar to the Z-scores determined for the template structures (TBSV,  $-5.18$ ; MNSV,  $-6.26$ ; CMV,  $-6.06$ ; TCV,  $-3.39$ ; Fig. 3). The MNSV virion map was downloaded from the VIPER database (viperd.b.scripps.edu/) and rendered using UCSF Chimera<sup>63</sup>.

**Phylogenetic analysis.** Multiple sequence alignments for RC-Rep and capsid genes were computed with MUSCLE<sup>64</sup> and manually curated (alignments are available from the authors upon request). Positions (560 and 289) were selected from the CP and RC-Rep alignments, respectively, and were used for subsequent phylogenetic analysis. Maximum-likelihood trees were calculated using FastTree<sup>65</sup> with Jones-Taylor-Thornton model of amino acid evolution and  $\gamma$ -CAT estimation of evolutionary rates across sites. Phylogenetic reconstructions with Bayesian MCMC (Markov chain Monte Carlo) yielded very similar tree topologies. The trees were annotated with ItoI<sup>66</sup>.

To test the monophyly of CHIV sequences in CP and Rep phylogenetic trees, two maximum-likelihood trees were computed for each alignment: one unconstrained and one with all CHIV sequences forced into a monophyletic group. TreeFinder<sup>67</sup> was used to compare the two topologies for each alignment through expected-likelihood weights and the approximately unbiased<sup>68</sup> test (Table 2).

**Estimation of evolutionary distances between proteins.** MEGA 5 (ref. 69) was used to assess evolutionary distances between protein sequences of capsid and RC-Rep genes (Jones-Taylor-Thornton model,  $\gamma$ -parameter set to the default value of 1.3). For ssDNA and ssRNA viruses, all available genomes were downloaded from NCBI, and clustered based on taxonomy (one genome for each species) and on global sequence similarity (threshold of 75% identity) with Uclust<sup>70</sup>. Comparisons were made within each taxonomic group (*Circoviridae*, *Geminiviridae*, *Nanoviridae* and *Tombusviridae*) and between CHIVs based on distinct multiple alignments computed with MUSCLE<sup>64</sup>. To keep the chart clear and viewable, only distances below 25 were taken, which removed 30 comparisons between *Geminiviridae* where RC-Rep protein distances were below 10 but capsid genes distances were between 25 and 100. The same set of sequences was used in the genome size box plot. Unclassified ssDNA and ssRNA viruses were not included in these analyses.

## References

- King, A. M. Q., Adams, M. J., Carstens, E. B. & Lefkowitz, E. J. *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses* (Elsevier Academic Press, 2011).
- Mochizuki, T. *et al.* Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome. *Proc. Natl Acad. Sci. USA* **109**, 13386–13391 (2012).
- Pietilä, M. K., Roine, E., Paulin, L., Kalkkinen, N. & Bamford, D. H. An ssDNA virus infecting archaea: a new lineage of viruses with a membrane envelope. *Mol. Microbiol.* **72**, 307–319 (2009).
- Yu, X. *et al.* A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. *Proc. Natl Acad. Sci. USA* **107**, 8387–8392 (2010).
- Senčilo, A., Paulin, L., Kellner, S., Helm, M. & Roine, E. Related haloarchaeal pleomorphic viruses contain different genome types. *Nucleic Acids Res.* **40**, 5523–5534 (2012).

6. Yamada, T. *et al.* New bacteriophages that infect the phytopathogen *Ralstonia solanacearum*. *Microbiology* **153**, 2630–2639 (2007).
7. Dunlap, D. S. *et al.* Molecular and microscopic evidence of viruses in marine copepods. *Proc. Natl Acad. Sci. USA* **110**, 1375–1380 (2013).
8. Nagasaki, K. *et al.* Previously unknown virus infects marine diatom. *Appl. Environ. Microbiol.* **71**, 3528–3535 (2005).
9. Roux, S., Krupovic, M., Poulet, A., Debroas, D. & Enault, F. Evolution and diversity of the *Microviridae* viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS One* **7**, e40418 (2012).
10. Rakonjac, J., Bennett, N. J., Spagnuolo, J., Gagic, D. & Russel, M. Filamentous bacteriophage: biology, phage display and nanotechnology applications. *Curr. Issues Mol. Biol.* **13**, 51–76 (2011).
11. Bennett, A., McKenna, T. & Agbandje-McKenna, M. A comparative analysis of the structural architecture of ssDNA viruses. *Comput. Math. Methods Med.* **9**, 183–196 (2008).
12. Krupovic, M. Recombination between RNA viruses and plasmids might have played a central role in the origin and evolution of small DNA viruses. *Bioessays* **34**, 867–870 (2012).
13. Krupovic, M. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. *Curr. Opin. Virol.* **3**, 578–586 (2013).
14. Ilyina, T. V. & Koonin, E. V. Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaeobacteria. *Nucleic Acids Res.* **20**, 3279–3285 (1992).
15. Koonin, E. V. & Ilyina, T. V. Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. *J. Gen. Virol.* **73**(Pt 10): 2763–2766 (1992).
16. Koonin, E. V. A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication. *Nucleic Acids Res.* **21**, 2541–2547 (1993).
17. Rosario, K., Duffy, S. & Breitbart, M. A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch. Virol.* **157**, 1851–1871 (2012).
18. Delwart, E. & Li, L. Rapidly expanding genetic diversity and host range of the *Circoviridae* viral family and other Rep encoding small circular ssDNA genomes. *Virus Res.* **164**, 114–121 (2012).
19. Martin, D. P. *et al.* Recombination in eukaryotic single stranded DNA viruses. *Viruses* **3**, 1699–1738 (2011).
20. Cherwa, J. E. & Fane, B. A. In *Encyclopedia of Life Sciences* (Wiley, Chichester, 2011).
21. Krupovic, M., Ravantti, J. J. & Bamford, D. H. Geminiviruses: a tale of a plasmid becoming a virus. *BMC Evol. Biol.* **9**, 112 (2009).
22. Dayaram, A. *et al.* Novel ssDNA virus recovered from estuarine Mollusc (*Amphibola crenata*) whose replication associated protein (Rep) shares similarities with Rep-like sequences of bacterial origin. *J. Gen. Virol.* **94**, 1104–1110 (2013).
23. Dolja, V. V. & Koonin, E. V. Common origins and host-dependent diversity of plant and animal viromes. *Curr. Opin. Virol.* **1**, 322–331 (2011).
24. Rosario, K., Duffy, S. & Breitbart, M. Diverse circovirus-like genome architectures revealed by environmental metagenomics. *J. Gen. Virol.* **90**, 2418–2424 (2009).
25. Rosario, K. *et al.* Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). *J. Gen. Virol.* **93**, 2668–2681 (2012).
26. Roux, S. *et al.* Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* **7**, e33641 (2012).
27. Ng, T. F. *et al.* High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. *J. Virol.* **86**, 12161–12175 (2012).
28. Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* **9**, 267–276 (2008).
29. Firth, C., Charleston, M. A., Duffy, S., Shapiro, B. & Holmes, E. C. Insights into the evolutionary history of an emerging livestock pathogen: porcine circovirus 2. *J. Virol.* **83**, 12813–12821 (2009).
30. Diemer, G. S. & Stedman, K. M. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. *Biol. Direct.* **7**, 13 (2012).
31. Stedman, K. Mechanisms for RNA capture by ssDNA viruses: grand theft RNA. *J. Mol. Evol.* **76**, 359–364 (2013).
32. Whon, T. W. *et al.* Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. *J. Virol.* **86**, 8221–8231 (2012).
33. Angly, F. E. *et al.* The marine viromes of four oceanic regions. *PLoS Biol.* **4**, e368 (2006).
34. Rosario, K., Nilsson, C., Lim, Y. W., Ruan, Y. & Breitbart, M. Metagenomic analysis of viruses in reclaimed water. *Environ. Microbiol.* **11**, 2806–2820 (2009).
35. Krupovic, M. & Forterre, P. *Microviridae* goes temperate: microvirus-related proviruses reside in the genomes of *Bacteroidetes*. *PLoS One* **6**, e19893 (2011).
36. Liu, H. *et al.* Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evol. Biol.* **11**, 276 (2011).
37. Vulot, D. *et al.* Metagenomes of the picoalga *Bathycoccus* from the Chile coastal upwelling. *PLoS One* **7**, e39648 (2012).
38. Habura, A., Hou, Y., Reilly, A. A. & Bowser, S. S. High-throughput sequencing of *Astrammnia rara*: sampling the giant genome of a giant foraminiferan protist. *BMC Genomics* **12**, 169 (2011).
39. Rochon, D., Lommel, S., Martelli, G. P., Rubino, L. & Russo, M. In *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*. (eds King, A. M. Q., Adams, M. J., Carstens, E. B. & Lefkowitz, E. J.) 1111–1138 (Elsevier Academic Press, 2011).
40. Koenig, R., Verhoeven, J. T., Fribourg, C. E., Pfeilstetter, E. & Lesemann, D. E. Evaluation of various species demarcation criteria in attempts to classify ten new tombusvirus isolates. *Arch. Virol.* **149**, 1733–1744 (2004).
41. Yokoi, T., Yamashita, S. & Hibi, T. The nucleotide sequence and genome organization of *Sclerophthora macrospora* virus A. *Virology* **311**, 394–399 (2003).
42. Heller-Dohmen, M., Gopfert, J. C., Pfanstiel, J. & Spring, O. The nucleotide sequence and genome organization of *Plasmopara halstedii* virus. *Virol. J.* **8**, 123 (2011).
43. Sherman, M. B. *et al.* Removal of divalent cations induces structural transitions in red clover necrotic mosaic virus, revealing a potential mechanism for RNA release. *J. Virol.* **80**, 10395–10406 (2006).
44. Koonin, E. V. & Dolja, V. V. A virocentric perspective on the evolution of life. *Curr. Opin. Virol.* **3**, 546–557 (2013).
45. Dijkeng, A., Kuzmickas, R., Anderson, N. G. & Spiro, D. J. Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS One* **4**, e7264 (2009).
46. Yutin, N. & Koonin, E. V. Hidden evolutionary complexity of nucleocytoplasmic large DNA viruses of eukaryotes. *Virol. J.* **9**, 161 (2012).
47. Pawlowski, J., Holzmann, M., Fahrni, J. F. & Hallock, P. Molecular identification of algal endosymbionts in large miliolid foraminifera: 1. Chlorophytes. *J. Eukaryot. Microbiol.* **48**, 362–367 (2001).
48. Santini, S. *et al.* Genome of *Phaeocystis globosa* virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. *Proc. Natl Acad. Sci. USA* **110**, 10800–10805 (2013).
49. Desnues, C. *et al.* Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc. Natl Acad. Sci. USA* **109**, 18078–18083 (2012).
50. Park, M. G., Yih, W. & Coats, D. W. Parasites and phytoplankton, with special emphasis on dinoflagellate infections. *J. Eukaryot. Microbiol.* **51**, 145–155 (2004).
51. Speir, J. A. & Johnson, J. E. Nucleic acid packaging in viruses. *Curr. Opin. Struct. Biol.* **22**, 65–71 (2012).
52. Koonin, E. V., Wolf, Y. I., Nagasaki, K. & Dolja, V. V. The complexity of the virus world. *Nat. Rev. Microbiol.* **7**, 250 (2009).
53. Krupovic, M., Prangishvili, D., Hendrix, R. W. & Bamford, D. H. Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. *Microbiol. Mol. Biol. Rev.* **75**, 610–635 (2011).
54. Lawrence, J. G., Hatfull, G. F. & Hendrix, R. W. Imbroglions of viral taxonomy: genetic exchange and failings of phenetic approaches. *J. Bacteriol.* **184**, 4891–4905 (2002).
55. Kristensen, D. M. *et al.* Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *J. Bacteriol.* **195**, 941–950 (2013).
56. Ignacio-Espinoza, J. C. & Sullivan, M. B. Phylogenomics of T4 cyanophages: lateral gene transfer in the ‘core’ and origins of host genes. *Environ. Microbiol.* **14**, 2113–2126 (2012).
57. Lavigne, R. *et al.* Classification of *Myoviridae* bacteriophages using protein sequence similarity. *BMC Microbiol.* **9**, 224 (2009).
58. Lavigne, R., Seto, D., Mahadevan, P., Ackermann, H. W. & Kropinski, A. M. Unifying classical and molecular taxonomic classification: analysis of the *Podoviridae* using BLASTP-based tools. *Res. Microbiol.* **159**, 406–414 (2008).
59. Krupovic, M. & Bamford, D. H. Order to the viral universe. *J. Virol.* **84**, 12476–12479 (2010).
60. Noguchi, H., Taniguchi, T. & Itoh, T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.* **15**, 387–396 (2008).
61. Marti-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325 (2000).
62. Wiederstein, M. & Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* **35**, W407–W410 (2007).
63. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
64. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
65. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
66. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475–W478 (2011).
67. Jobb, G., von Haeseler, A. & Strimmer, K. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol. Biol.* **4**, 18 (2004).

68. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).
69. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
70. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
71. Hewson, I. *et al.* Metagenomic identification, seasonal dynamics and potential transmission mechanisms of a *Daphnia*-associated putative RNA–DNA hybrid virus in two temperate lakes. *Limnol. Oceanogr.* **58**, 1605–1620 (2013).

### Acknowledgements

This work was supported by the French national programme Ecosphère continentale et côtière (EC2CO), project CAVIAR (CommunAutés de Virus à ARN). S.R. was supported by a PhD grant from the French defence procurement agency (DGA, Direction Générale de l'Armement). D.V. was supported by PHYTOMETAGENE (JST-CNRS), METAPICO (Genoscope) and Micro B3 (funded by the European Union, contract 287589).

### Author contributions

S.R. and M.K. designed and performed the research, analysed the data and wrote the paper; F.E. and G.B. analysed the data and contributed new reagents/analytic tools; D.V. and P.F. contributed new reagents/analytic tools.

### Additional information

**Accession Codes:** All CHIV genome sequences assembled in this study are available as annotated GenBank-formatted files through Dryad Digital Repository (datadryad.org), doi: 10.5061/dryad.19m2k.

**Supplementary Information** accompanies this paper at <http://www.nature.com/naturecommunications>

**Competing financial interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**How to cite this article:** Roux, S. *et al.* Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. *Nat. Commun.* **4**:2700 doi: 10.1038/ncomms3700 (2013).