ARTICLE

Repeated species radiations in the recent evolution of the key marine phytoplankton lineage Gephyrocapsa

El Mahdi Bendif1, Bruno Nevado1, Edgar L.Y. Wong1, Kyoko Hagino2, Ian Probert3, Jeremy R. Young4, Rosalind E.M. Rickaby5 & Dmitry A. Filatov1

Phytoplankton account for nearly half of global primary productivity and strongly affect the global carbon cycle, yet little is known about the forces that drive the evolution of these keystone microscopic organisms. Here we combine morphometric data from the fossil record of the ubiquitous coccolithophore genus Gephyrocapsa with genomic analyses of extant species to assess the genetic processes underlying Pleistocene palaeontological patterns. We demonstrate that all modern diversity in Gephyrocapsa (including Emiliania huxleyi) originated in a rapid species radiation during the last 0.6 Ma, coincident with the latest of the three pulses of Gephyrocapsa diversification and extinction documented in the fossil record. Our evolutionary genetic analyses indicate that new species in this genus have formed in sympatry or parapatry, with occasional hybridisation between species. This sheds light on the mode of speciation during evolutionary radiation of marine phytoplankton and provides a model of how new plankton species form.

1 Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK. 2 Center for Advanced Marine Core Research, Kochi University, Nankoku, Kochi 783-8502, Japan. 3 Sorbonne Université - CNRS, Roscoff Culture Collection, FR2424 Station Biologique de Roscoff, 29680 Roscoff, France. 4 Department of Earth Sciences, University College London, London WC1E 6BS, UK. 5 Department of Earth Sciences, University of Oxford, Oxford OX1 3AN, UK. Correspondence and requests for materials should be addressed to D.A.F. (email: Dmitry.Filatov@plants.ox.ac.uk)
Marine phytoplankton play an important ecological role underpinning food webs in the ocean and are responsible for about half of global primary productivity. Surprisingly, however, little is known about how marine plankton species originate and evolve. Adaptation and speciation processes may work in rather different ways in relatively small subdivided populations of terrestrial organisms and globally ubiquitous populations of abundant marine plankton. In particular, it is unclear how new plankton species form in relatively homogenous habitats, such as in the open ocean, where no physical barriers to gene flow exist to promote allopatric speciation (i.e. the passive divergence of isolated populations) that is considered to be the most common speciation scenario in terrestrial organisms. Ecological reasons for the 'paradox of the plankton', the unexpected diversity of plankton species in a seemingly homogenous environment, have been analysed in multiple studies, but still remain obscure in terms of evolutionary genetic processes.

The calcifying coccolithophores (Haptophyta) represent an excellent model group to study the evolutionary processes underlying plankton speciation by integrating genomic, biological, biogeographic and palaeontological data. This group comprises around 200 well-described extant species which are widely distributed in modern oceans. Coccolithophores form calcium carbonate scales, called coccoliths, which have complex and distinct architectures, allowing identification of morphospecies in extant diversity and the fossil record. Since their appearance around 220 Ma, coccoliths have formed massive calcareous deposits (e.g. the White Cliffs of Dover in southeast England) that serve as a sedimentary buffer of ocean chemistry and a major long-term carbon store that has significantly affected the global carbon cycle and Earth climate. Marine pelagic sediments provide an enormously abundant and essentially continuous fossil record of coccoliths, and this has been extensively studied through the Quaternary, which is consistent with similar size changes in their relatives Reticulofenestra throughout the Cenozoic. The gradual size increase may be an instance of the fl 20. The gradual size increase may be an instance of the gradual increase in organism size within a lineage with evolutionary time, which was first described for terrestrial animals, but has also been reported in the marine realm. Similar patterns of size shift over time have been reported in other marine protist groups, including diatoms, dinoflagellates and planktic foraminifera. The pattern is particularly well-marked in the genus *Gephyrocapsa*, though the evolutionary processes underpinning the pulses in coccolith size change remain unresolved.

Here we combine stratophenetic datasets on the *Gephyrocapsa* fossil record with genome sequence data from a diverse set of 10 strains isolated across the world oceans to integrate evolutionary analysis of detailed fossil records and extant patterns of genetic variation. Based on whole-genome sequences, we reconstruct the phylogeny for five closely-related Noëlaerhabdaceae species and infer their evolutionary history and likely mode of speciation. The combined analysis of fossil record and genome sequence data reveals evidence for repeated species radiations and extinctions in this lineage through the Quaternary.

Results and discussion

**Genome sequencing resolves *Gephyrocapsa* phylogeny.** Our analysis included the type species of the *Gephyrocapsa* genus, *G. oceanica*, that is widespread (Fig. 1a) and relatively straightforward to isolate and culture, as well as three other *Gephyrocapsa* species (*G. muellerae, G. parvula* and *G. ericsonii*; Fig. 1b), that were only recently isolated via a high-throughput sorting approach. The analysis also included the most widely distributed and abundant coccolithophore species in the modern oceans, *Emiliania huxleyi* (Fig. 1b), that is very closely related to the *Gephyrocapsa* genus and, as we show below, should be included in this genus. Previous attempts to resolve the phylogenetic relationships between members of *Gephyrocapsa* and *E. huxleyi* resulted in rather ambiguous patterns demonstrating conspicuous conflicts between molecular and interpretations based on morphological and palaeontological evidence. In order to resolve phylogenetic relationships and analyze evolutionary processes driving speciation in this ecologically important phytoplankton genus, we sequenced the whole genomes of 8 *Gephyrocapsa* strains isolated from across the world oceans and combined these data with 2 other publicly available genomes (Supplementary Table 1). Based on morphological and previous studies, these 10 *Gephyrocapsa* strains were classified as belonging to five different species (Fig. 1b and Supplementary Fig. 1), covering almost the full range of extant morphological diversity in this genus.

Sequence alignments based on whole-genome sequences were used to reconstruct fully resolved phylogenies for five *Gephyrocapsa* species. Regardless of the phylogenetic reconstruction method used, topologies of resulting phylogenies were identical and all nodes in the trees were highly supported (100% bootstrap; Fig. 2a) and all nodes in the trees were highly supported (100% bootstrap; Fig. 2a and Supplementary Figs. 2-4). The consensus phylogeny reconstructed from genome sequence data closely matches morphological taxonomy, with nearly all strains of the same morphospecies clustering together. This indicates that morphology-based definitions of *Gephyrocapsa* species (as in Young et al.) correspond to biological species. This is consistent with a previous analysis of multiple *E. huxleyi* strains sampled across the world oceans that demonstrated genetic cohesiveness of this species. The only mismatch of the genome-based phylogeny with morphology-based definition of species was the clustering of *G. parvula* strain RCC4033 with *G. ericsonii* strains RCC4032 and RCC4033 in the shallow *parvula/ericsonii* clade. However, even this result is supported by morphological evidence since these two species, in fact, represent a complex of intergrading morphotypes that frequently co-occur in the water column and in the sediments of eutrophic tropical oceans. Based on the very small genetic distance between *G. parvula* and *G. ericsonii*, compared to other distances in the phylogeny (Fig. 2a), we suggest that these two morphospecies are likely to be conspecific. Consistent with the previous studies, our phylogenomic results support the use of the species name *Gephyrocapsa huxleyi* instead of the commonly used *Emiliania huxleyi* as this lineage is nested within the *Gephyrocapsa* genus (Fig. 2a). Hence, we hereafter use *G. huxleyi* instead of *E. huxleyi*.

Despite the clear clustering of the *Gephyrocapsa* strains by morphospecies and the 100% bootstrap support for all nodes in the tree (Fig. 2a), a considerable proportion (11.6%) of individual gene trees reconstructed from different 10 kb parts of the genome showed incongruence with the species tree in the densiTREE plot (Fig. 2b). In particular, Shimodaira–Hasegawa tests showed that
11.6% of the 6278 trees analysed, significantly ($P < 0.05$) rejected the species tree. This type of incongruence can be caused by phenomena, such as incomplete lineage sorting (ILS) and is very common in genomes of different organisms. The extent of phylogenetic incongruence caused by ILS is proportional to the effective population size of the ancestral species and can, therefore, be used to estimate the population size of species that are extinct. This allowed us to study the demographic history of speciation throughout the Gephyrocapsa phylogeny.

How do new plankton species evolve. The genome sequence data for 10 Gephyrocapsa strains allowed us to study evolutionary genetic processes underpinning speciation in this genus. For this purpose, we used Bayesian inference under the multispecies coalescent model that estimates effective population sizes ($N_e$) at every node in the phylogeny while taking into account the possibility of interspecific gene flow. The results of this analysis are summarised in Fig. 2c, with the sizes of the circles proportional to estimated population sizes of extant and ancestral Gephyrocapsa species. Importantly, these genetic diversity-based estimates of Gephyrocapsa population sizes for extant species are congruent with current species abundance. In particular, G. huxleyi, which has been globally abundant since the beginning of the Holocene and became as abundant as other Gephyrocapsa species 70 ka, is now the most common coccolithophore in modern oceans (with a population size of the order of $10^{22}$ cells), which is consistent with the large population size inferred for this species compared to other extant Gephyrocapsa (Fig. 2c; Supplementary Fig. 5 and Supplementary Table 2). Furthermore, the fossil-based estimates of absolute and relative species abundance in this genus through time (Fig. 2d, e) are broadly congruent with our evolutionary genetic inference of population sizes in the ancestral Gephyrocapsa species (Fig. 2c). In particular, the common ancestor of all extant Gephyrocapsa species (node 1) and the common ancestor of all Gephyrocapsa species except G. oceanica (node 2) had the largest population sizes, which is congruent with particularly high Gephyrocapsa coccolith counts in sediments around 550–300 Ka (Fig. 2e). The origin of G. huxleyi around 300 Ka, that is well documented in the fossil record, apparently occurred in a very large population (node 2; Fig. 2c) without a significant speciation bottleneck (i.e. without a period of reduced population size due to origination of a species in a very small population). The highly distinctive phenotype of G. huxleyi (Fig. 1b) likely evolved throughout the range of the ancestral Gephyrocapsa species, as suggested by global synchrony of this event in the sediment record.

Our analyses indicate that interspecific hybridisation and low intensity gene flow has commonly occurred between Gephyrocapsa species throughout their evolutionary history. The estimates of interspecific gene flow, based on Bayesian inference under the multispecies coalescent model, are summarised in Fig. 2c as arrows, revealing gene flow between most nodes in the phylogeny. This is consistent with results of a Patterson’s $D$-statistic test that detected gene flow between G. huxleyi and G. muellerae and between G. huxleyi and G. ericsonii/parvula (Supplementary Fig. 6; Supplementary Tables 3 and 4). Thus, closely-related Gephyrocapsa species are not completely reproductively isolated and they co-occur to occasionally form hybrids. This finding rules out a purely allopatric mode of speciation and
indicates that new Gephyrocapsa species formed in parapatry or sympatry, which is consistent with the frequent co-occurrence of different Gephyrocapsa species in the water column and in marine sediments.

Repeated species radiations in Gephyrocapsa fossil record. Using morphometric data from Matsuoka & Okada, we analysed the evolution of Gephyrocapsa coccolith morphology over the Pleistocene (1.8 Ma; Fig. 3a) measuring coccolith size, bridge angle and central opening (Fig. 3b; Supplementary Table 5) in successive populations with a sampling resolution of about 40 ka. The study revealed 3 pulses of coccolith size increase punctuated by episodes of size variance reduction. For convenience, we refer to these events of increasing size of coccoliths as ‘Matsuoka–Okada cycles’ (MO-cycles), and label them from oldest to most recent as MO1, MO2 and MO3 (Fig. 3a, Supplementary Figs. 7 and 8). Before the Pleistocene, Gephyrocapsa were typically rare and small (coccoliths <3 µm) and they become more abundant and larger (coccoliths >4 µm) from about 1.7–1.2 Ma (MO1, Fig. 3a). This episode of coccolith size increase at 1.7 Ma was followed by a gradational process whereby distinct population clusters were recognised, with the two larger forms (species A and B in Matsuoka & Okada’s Fig. 3a) disappearing at the same time. Only small Gephyrocapsa (<3 µm) were then present for about 250 ka (about 1.2–0.95 Ma). This event has been used for over 40 years as a key datum in biostratigraphic studies including ODP and IODP scientific ocean drilling projects. This cycle was followed by two more sequences of gradual size increase (MO2 and MO3, Fig. 3a), initiated by an abrupt increase around 950 ka and separated by a shorter interval of smaller specimens only, apparent through the reduction of coccolith size variance (top-down, bottom-up) at around 0.5 ka. The available evidence indicates that most size reduction events were abrupt and globally synchronous, whilst size increases were gradational (Fig. 3a).

MO1 to MO3 events are marked by a synchronous increase in the abundance of both large and small Gephyrocapsa species (Fig. 2d), indicating a turnover in the assemblage, rather than merely fluctuations in relative abundance of the same large- and small-sized species. In order to test this, we estimated the ages of
speciation events in the Gephyrocapsa genus from our newly generated genome sequence data. We used a relaxed molecular clock with the calibration point corresponding to the most reliable fossil date, the first appearance of G. huxleyi around 300 ka (node 2 on Fig. 2a). This allowed us to date other nodes in the phylogeny, revealing that the most recent common ancestor (MRCA) of extant Gephyrocapsa species arose around 553 ka (node 1 on Fig. 2a and Supplementary Table 6). If MO-cycles were caused by the change in abundance of the same large- and small-sized species through time, then the MRCA of extant Gephyrocapsa species would be expected to be several million years old (coincident with first appearance of the Gephyrocapsa ~3.5 Ma). On the other hand, if each MO-cycle represents an independent burst of species diversification into larger and smaller forms, followed by extinction, then we would expect the MRCA to be much younger—around 0.5 Ma—the age of the last MO-cycle. Our estimate of the Gephyrocapsa MRCA age ~553 ka clearly demonstrates that all extant diversity in this genus evolved within the most recent MO-cycle 3, which indicates that MO-cycles likely represent consecutive species radiations, followed by globally synchronous extinctions. This is also supported by the detailed analysis of coccolith morphology. In particular, the comparison of larger coccoliths from different MO-cycles revealed that they consistently differed from each other in characteristics such as bridge angle and central opening size (Supplementary Fig. 1 and Supplementary Table 7), indicating that different MO-cycles were caused by the spread of genetically distinct species: consecutively G. lumina Burky, G. omega Burky and G. oceanica Kamptner.

Evolution of extant Gephyrocapsa diversity within the last MO-cycle 3 involved significant morphological diversification (Figs. 1b and 3a), including evolution of a wide range of coccolith sizes and degrees of calcification and at least two independent events of bridge loss, one corresponding to the establishment of G. huxleyi and the other to G. parvula (Fig. 1b; Supplementary Fig. 9). This resolves the ambiguity of whether the bridge was gained or lost through time and illustrates the evolutionary dynamic of this diagnostic character. This evolutionary pattern implies that more non-bridged members of the Noëlaerhabdaceae might have derived from bridged forms throughout the evolution of Gephyrocapsa, with further instances of bridge loss likely having occurred in previous MO-cycles. Furthermore, the example of G. ericsonii and G. parvula also suggests that other co-occurring bridged and non-bridged Gephyrocapsa could have been intergrading in the past.

Within each MO-cycle, there is a pattern of gradual coccolith size increase, followed by abrupt decrease. It has been suggested that such gradual increase in size, often referred to as Cope’s rule, is a passive phenomenon caused by a lower physiological limit in size. Therefore, after each extinction event, when primarily the smaller less specialized taxa survive, the maximal size in a group gradually drifts upwards, while the...
periods of increased amplitude in oceanic duration and pacing of the MO events seem to coincide with impact abiotic and biotic processes\textsuperscript{48}. In the open ocean and the Milankovitch glacial cycles of the Pleistocene (Fig. 3b), but the conditions\textsuperscript{24,43}. Temperature possible drivers of repeated plankton species radiations observations indicate that pulsed events in plankton evolution growth, while modifying available niches\textsuperscript{49}. In removing the opportunistic responses of closely-related co-occurring species, the increase in size could not be explained by neutral ancestry, with characteristics similar to MO-cycles, have also been identified, such as periods of dramatic size reduction\textsuperscript{13}. Furthermore, the history of the coccolithophore genus Calcidiscus displays the repeated development of coccoliths similar in appearance but at stratigraphically distant intervals\textsuperscript{41}. These observations indicate that pulsed events in plankton evolution may be widespread across the coccolithophores and our results indicate that they represent species radiations separated by abrupt extinctions.

Possible drivers of repeated plankton species radiations. MO-like macroevolutionary patterns in a range of plankton groups have been inferred to relate to changes in environmental conditions\textsuperscript{24,43}. Temperature fluctuations are likely to selectively impact abiotic and biotic processes\textsuperscript{48}. In the open ocean and the photic zone, variation in the thermocline depth and/or strength is well known to affect phytoplankton community structure through changes in the availability of light and nutrients for growth, while modifying available niches\textsuperscript{49}. In removing the dominance of one keystone species, climate change may promote opportunistic responses of closely-related co-occurring species bearing relative adaptive advantages\textsuperscript{50}.

MO events (Fig. 3a) are too extended to follow the periodic Milankovitch glacial cycles of the Pleistocene (Fig. 3b), but the duration and pacing of the MO events seem to coincide with phases of different maximum global ice volume expressed as periods of increased amplitude in oceanic δ\textsuperscript{18}O\textsuperscript{51}, and alkenone-derived SST\textsuperscript{52} cycles towards the modern day (Supplementary Fig. 8). Northern hemisphere glaciation intensified ~2.58 Ma\textsuperscript{53}, long before MO1 cycle, and these ice sheets oscillated at a regular 41 kyr period spanning from 2.58 to ~0.95 Ma (Fig. 3b). MO1 started after a global cooling event in SST around 1.8 Ma (event occurring from 2.1 to 1.8 Ma\textsuperscript{34,55}) and ended abruptly ~1.2 Ma with the onset of the Mid-Pleistocene climate Transition (MPT; ~1.2 Ma to ~0.50 Ma when climate evolved from being dominated by a 41 ka to a 100 ka periodicity\textsuperscript{56}). There was then a period of ~250 kyr of apparent stasis in coccolith size followed by a documented abrupt increase in ice volume and increased ocean storage of carbon\textsuperscript{31,57} around 950 ka, which coincided with the upturn in coccolith size during the MO2 cycle. The subsequent extinction and reduction in coccolith size of the MO2 cycle coincided with the end of the MPT ~500 ka when the glacial cycles in climate first attained their greatest amplitude and a regular 100 kyr periodicity. The extinction that triggered the start of MO3 also coincided with the most northerly migration of the subtropical front to impinge on southern hemisphere continents\textsuperscript{58} and potentially restrict flow of low latitude waters between ocean basins. If oceanic fronts do act as barriers to gene flow\textsuperscript{59-61}, then this restriction could have isolated the low latitude basins making endemic populations prone to higher extinction rates. MO3 then proceeded during the regular 100 kyr fluctuations.

Other plankton also appear to evolve in response to step changes in global cooling. Evolution of diatoms in the Southern Ocean has shown diversification rates punctuated by several instances of massive extinctions closely related to global stepwise cooling suggesting a sensitivity to long-term climatic events\textsuperscript{62}. Beyond the marine realm, the evolution of plants and animals also show climatically-sensitive evolutionary trends that are well characterised in hominids, Bovidae, birds and plants\textsuperscript{63}. Although the mechanisms may be different in different realms, the evolutionary pulsed turnover of new species\textsuperscript{63} may generally be associated with step changes in environmental conditions.

Concluding remarks. This study is the first to combine the analysis of plankton fossil data with evolutionary genetic analysis of whole-genome sequences to advance our understanding of plankton speciation. Our study indicated that pulses of coccolith (and likely cell\textsuperscript{17,64}) size change observed in the plankton fossil record are caused by repeated species radiations rather than fluctuations in the relatively abundant large- and small-celled species. These species radiations comprise rapid diversification of coccolith sizes and morphology, followed by abrupt globally synchronous extinctions recorded in ocean sediments every ~0.5 million years. Similar patterns of cycles in size appear to be common in many plankton groups\textsuperscript{13} and may also represent repeated pulses of species radiations and extinctions that lead to high turnover in diversity. Our evolutionary genetic analyses of coccolithophore genomes reveal the mode of speciation during species radiations of marine phytoplankton. In particular, our results rule out purely allopatrial evolution in *Gephyrocapsa* and indicate that new species in this genus have formed in sympatry or parapatry, with occasional hybridisation between the species.

These results help to build a link between macroevolutionary patterns observed in the fossil record and micro-evolutionary processes underlying new species formation in marine microplankton. The combined analysis of fossil and evolutionary genomic sequence data is a major step forward from fossil/sediment-only or sequence-based only studies. Further work with continuous records of sufficient resolution combined with evolutionary genomic sequence data, could resolve whether pulsed radiations are a fundamental characteristic of macro-evolution both terrestrial and marine. Establishing the prevalence of such cycles is the first step to understanding the macroevolutionary process. The integration of population and evolutionary genetics with diverse fossil and climatic records could reveal much further insight into the mechanistic coevolution of biodiversity, productivity, the carbon cycle and climate.

Methods

**Origin and morphological characterisation of strains.** Clonal Noëlaerhabdaceae strains (Supplementary Table 1) from the Roscoff Culture Collection (RCC; roscoff-culture-collection.org) were maintained in K/2(-Si,-Tris,-Cu) medium\textsuperscript{65} at 17 °C with 50 µmol-photons.m\textsuperscript{−2} s\textsuperscript{−1} illumination provided by daylight neon tubes with a 14:10 h L:D cycle. Cultured cells were harvested at early exponential growth phase on 0.22 µm nucleopore filters, then dried in a 55 °C oven for 2 h. Following gold coating, filters were observed with a Phenom scanning electron microscope. Morphometric measurements (Supplementary Table 5) were carried out according to Bollmann\textsuperscript{66} with a minimum of 60 isolated coccoliths per strain using ImageJ software (http://imagej.nih.gov/ij/).

**Stratophenetic data.** Our reconstruction of the fossil record of *Gephyrocapsa* is based on the study of Okada and Matsuoka\textsuperscript{67}. Morphometric data from this study, including measurements of coccolith size, bridge angle and central opening
DNA extraction. Cell cultures were harvested by centrifugation (4500 g x 15 min), washed twice with TE buffer, and suspended in 10 mL of lysis buffer (Tris, 0.1 M; EDTA, 0.05 M; NaCl, 0.1 M; 1% SDS; 2% Na-lauroylsarcosine, proteinase K 200 mg/mL; pH 8.0) and incubated at 55 °C for 2 h for extraction of total genomic DNA. DNA was then purified with equal volumes of phenol and chloroform and precipitated with ethanol. For each sample, quantifications of nucleic acids were performed either with a Qubit 3.0 fluorometer (ThermoFisher Scientific, Inc.) or a Nanodrop. DNA extracts were sent to the Wellcome Trust Centre for Human Genomics, Oxford (WTCHG) for sequencing. Paired-end libraries were prepared individually, barcoded, and then combined prior to sequencing. Libraries were sequenced using an Illumina HiSeq 2500 sequencing platform to produce 150 base-pair (bp) paired-end reads. The amount of raw data generated for each strain is individually, barcoded, and then combined prior to sequencing. Libraries were sequenced using an Illumina HiSeq 2500 sequencing platform to produce 150 base-pair (bp) paired-end reads. The amount of raw data generated for each strain is

Mapping of reads. After quality trimming with Trimmomatic, the sequence reads from *Gephyrocapsa* strains were mapped to the *G. huxleyi* CCMPP1516 reference genome with BWA-MEM. Despite low sequence divergence (<3% total sites on average between strains analysed), the proportion of reads mapped to reference was relatively low (40–69%; Supplementary Table 1) likely due to the known variability of *G. huxleyi* pan-genome with a large proportion of the genome missing in many strains. Duplicated reads were removed using Picard (https://broadinstitute.github.io/picard/). The Genome Analyses Toolkit (GATK) was then used to base-call recalibration, local remapping around indels (insertions/deletions), and SNPs (Single Nucleotide Polymorphisms) calling. For downstream analyses, obtained SNPs in vcf file format were converted into fasta multiple alignment sequences per contig (7795 contigs) using a custom script vcf2fas (available at https://github.com/brunonevado/vcf2fas).

Phylogenetic inference. For phylogenetic analysis of the Noelaerhabdaceae, alignments of each contig with at least one missing individual were discarded reducing the length of the alignment from 167 to 157 megabases, accounting for 4,015,458 positions which were phylogenetically informative. Initially, we conducted a species tree reconstruction using a concatenation-based approach, by first concatenating all 2137 contigs. We then performed a maximum likelihood (ML) tree inference on this concatenated alignment in RAxML 8.2 using the GTRGAMMA model and 100 bootstrap replicates. This approach allows production of a bootstrapped species tree with branch lengths. Next, we reconstructed a species phylogeny using a multispecies coalescent-based approach to account for incomplete lineage sorting (ILS). Contig-alignments were split in 13,810 separate regions 10 kb long, of which 6278 were used for the phylogenetic analysis after excluding the alignments with >20% of gaps or missing data. For each of these 6,278 alignments, we performed a phylogenetic reconstruction using the GTRGAMMA model and 100 bootstrap replicates in RAxML. Best ML trees with bootstrap replicates were then used to produce a species tree using ASTRAL. ASTRAL 100 bootstrap replicates using both site-wise and gene-wise resampling. The same ASTRAL analysis was repeated with shorter (5 kb long) windows, which yielded the same species tree as the analysis with 10 kb long windows (Supplementary Fig. 4).

An Isochrysis tulea tree was reconstructed in order to provide a root to the Noelaerhabdaceae phylogeny by using two outgroups: *Isochrysis lutea* and *Isochrysis galbana*, which are the closest available relatives of *Gephyrocapsa*. The genome sequence for *T. tulea* strain CCAP226/14 was obtained from http://www.seanoe.org/data/00361/471717/4, while the transcriptome sequence of *Isochrysis galbana* strain CCMPP1323 was obtained from the Marine Microbial Transcriptome Sequencing Project. To identify orthologs we extracted coding sequences (CDS) from the published *G. huxleyi* genome and compared them with the orthogroup sequences using Orthofinder 8.2.2 with standard parameters. The corresponding regions were extracted from *Gephyrocapsa* alignments, translated to predicted protein sequence and aligned with orthogroup sequences using MAFFT 8. The resulting alignments were filtered with alignment columns with >20% missing data excluded before concatenating all genes. We then performed a maximum likelihood tree inference on this concatenated alignment with RAxML using the PROTGAMMAAUTO model and 100 bootstrap replicates. To assess the reliability of the position of the root we also used the more complex mixture model CAT-GTR. The position of the root for the Noelaerhabdaceae phylogeny on the branch connecting *G. oceanica* with other *Gephyrocapsa* species was uncertain from the orthogroup and model used (Supplementary Figs. 2 and 3).

Molecular clock analyses. In order to better interpret patterns of speciation and morphological changes, we estimated divergence times for the *Gephyrocapsa* using uncorrelated and autocorrelated relaxed-clock approaches implemented in the mcmctree program 41 from the PAML 4.9b package 8. As mcmctree is able to infer the species tree, we used this analysis on the 6278 genomic sequence alignments 10 kb long obtained as described above. Alignment columns with missing data, gaps and ambiguity characters were excluded from analysis (mcmctree setting cleandata = 0). A uniform prior with min = 290 Ka and max = 350 Ka was used as a calibration time for the divergence between *G. huxleyi* and other *Gephyrocapsa* (node 2 on Fig. 2a, based on first occurrence of *G. huxleyi* in the fossil record 291 Ka, which is the only confident calibration point available in the fossil record of the group. As the mcmctree program requires definition of a “safe maximum” for the age of the root, we set this parameter to “-10” Ma. The analysis was conducted with an ultrametric clock model and the default the G5 model with full likelihood option (data = 1) in mcmctree and comprised two runs of 500,000 MCMC steps after a 50,000 burn-in. To test the robustness of the age estimates to the parameter values and the type of relaxed molecular clock used, we also conducted multiple shorter (200,000 MCMC steps after a 20,000 burn-in) runs with a subset of the 6278 genomic sequence alignments that included 500 randomly selected alignments. This subset was used for running the mcmctree program with both the independent rates model (clock = 2) and the correlated rates model (clock = 3) under HKY85 + G5 model (model = 4; alpha = 0.5) and full likelihood option (data = 1) in mcmctree and comprised two runs of 500,000 MCMC steps after a 50,000 burn-in. To test the robustness of the age estimates to the parameter values and the type of relaxed molecular clock used, we also conducted multiple shorter (200,000 MCMC steps after a 20,000 burn-in) runs with a subset of the 6278 genomic sequence alignments that included 500 randomly selected alignments. This subset was used for running the mcmctree program with both the independent rates model (clock = 2) and the correlated rates model (clock = 3) under HKY85 + G5 model with full likelihood option, which yielded similar results (Supplementary Table 6). Posterior distributions were then generated for each branch length in this analysis.

To determine how much of the observed variation among the 10 kb fragment trees was due to genuine incongruence, rather than simply lack of phylogenetic signal, we used Shimodaira–Hasegawa (SH) tests 83 implemented in CONSEL version 0.2 on all 10 kb alignments. The procedure first uses the phylogenetic inference program PhyML version 3.084 over two runs for each locus. The first run uses an unconstrained topology, and the second run constrains the topology to that of the species tree. Both runs were performed using the GTR substitution model and 100 bootstrap replicates. Site-likelihoods from these runs were then compared using the Shimodaira–Hasegawa exact test with 10,000 permutations. The expected number of incongruent sites in the region was calculated as the number of sites in the tree where branch lengths are averaged across all trees for a given topology with the following settings (star tree, consensus width = 1, consensus intensity 28.1, and default values for all other settings).

To analyse how much of the observed variation among the 10 kb fragment trees was due to genuine incongruence, rather than simply lack of phylogenetic signal, we used Shimodaira–Hasegawa (SH) tests 83 implemented in CONSEL version 0.2 on all 10 kb alignments. The procedure first uses the phylogenetic inference program PhyML version 3.084 over two runs for each locus. The first run uses an unconstrained topology, and the second run constrains the topology to that of the species tree. Both runs were performed using the GTR substitution model and 100 bootstrap replicates. Site-likelihoods from these runs were then compared using the Shimodaira–Hasegawa exact test with 10,000 permutations. The expected number of incongruent sites in the region was calculated as the number of sites in the tree where branch lengths are averaged across all trees for a given topology with the following settings (star tree, consensus width = 1, consensus intensity 28.1, and default values for all other settings).
the topology: (((P1,P2),P3),Outgroup). If the incongruence is due to ILS, the frequencies of these site patterns are expected to be equal, but in the case of introgression between P3 and either P1 or P2, they are expected to be biased toward the site pattern that clusters the introgressed taxa together. Block Jackknifing (with each locus representing a single block in the context of our dataset) was then used to determine significance. We used a custom ABA/BABA script (available at https://github.com/brunonevado/calcD_from_fas) to test every dataset. The custom programs and scripts are available at GitHub: https://github.com/brunonevado/.

Data availability

All genome sequencing data generated in this study have been deposited in the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/) and are accessible under bioproject number PRJNA532411. Two previously published G. huxleyi genotypes (ACC030) that were used in this study are available from NCBI under accession numbers ERR695588 and ERR695590. Coccolith morphometric data accessible under bioproject number PRJNA532411. Two previously published genotypes of Emiliania huxleyi (Haptophyta) were used in this study. Emiliania huxleyi (Haptophyta) are expected to be biased introgression between P3 and either P1 or P2, they are expected to be biased introgression between P3 and either P1 or P2, they are expected to be biased introgression between P3 and either P1 or P2, they are expected to be biased the paradoxis of the plankton. The paradox of the plankton. Nature 281, 237–240 (1986).

2. Rengefors, K., Kremp, A., Reusch, T. B. H. & Wood, A. M. Genetic diversity and evolution in eukaryotic phytoplankton: revelations from population genetic studies. J. Plankton Res. 39, 165–179 (2017).

3. De Vargas, C., Aubry, M. P., Probert, I. & Young, J. in Evolution of Primary Producers in the Sea (eds P. G. Falkowski & A. H. Knoll) Ch. 12, 251–285 (Elsevier, 2007).

4. de Vargas, C. et al. Eukaryotic plankton diversity in the sunlit ocean. Science 348, 1261605 (2015).

5. Abbott, R. et al. Hybridization and speciation. J. Evolution Biol. 26, 229–246 (2013).

6. Coyne, J. A. & Orr, H. A. Speciation (Sinauer Associates, 2004).

7. Hutchinson, G. E. The paradox of the plankton. Am. Naturalist 95, 137–145 (1961).

8. Scolesh, N., Hegreness, M. & Kishony, R. Evolution exacerbates the paradox of the plankton. Proc. Natl Acad. Sci. USA 105, 12365–12369 (2008).

9. Benner, I. et al. Emiliania huxleyi increases calcification but not expression of calcification-related genes in long-term exposure to elevated temperature and pCO2. Philos. Trans. R. Soc. Lond. B Biol. Sci. 368, 20130049 (2013).

10. Young, R. R. et al. A guide to extinct coccolithophore taxonomy. J. Nannoplankton Res. 125 (2003).

11. Westbroek, P. et al. A model system approach to biological climate forcing. Ex. Emiliania huxleyi. Glob. Planet Change 28, 47–46 (1993).

12. Erba, F. The first 150 million years history of calcareous nannoplankton: biosphere–geosphere interactions. Palaeogeogr. Palaeocl. 232, 237–250 (2006).

13. Bown, P. R. Calcareous nannofossil biostratigraphy. (Springer Netherlands, 1998).

14. Backman, J., Raffi, L., Rio, D., Fornaciari, E. & Pälike, H. Biozonation and biochronology of Miocene through Pleistocene calcareous nannofossils from the middle Miocene. Nat. Comm. 7, 10284 (2016).

15. Reusch, T. B. & Wood, A. M. Genetic diversity and evolution in eukaryotic phytoplankton: revelations from population genetic studies. Ann. Rev. Ecol. Evol. Syst. 35, 467–490 (2004).

16. Hagen, K. & Young, J. in Marine protists: diversity and dynamics (eds Ohnishi Ohsukata et al.) 311–330 (Springer Japan, Tokyo, 2015).

17. Smith, D. R. & Keeling, P. J. Twenty-fold difference in evolutionary rates between the mitochondrial and plastid genomes of species with secondary red plastids. J. Eukaryot. Microbiol. 59, 181–184 (2012).

18. Raffi, I. et al. A review of calcaro nannofossil astrochronology encompassing the past 25 million years. Quat. Sci. Rev. 25, 3133–3137 (2006).

19. Liebeholz, A., Koening, W. D. & Bjørnstad, O. N. Spatial synchrony in population dynamics. Ann. Rev. Ecol. Evol. Syst. 36, 467–490 (2004).

20. Bendik, E. M., Probert, I., Young, J. R. & von Dassow, P. Morphological and phylogenetic characterization of new Gephyrocapsa isolates suggests introgressive hybridization in the Emiliania/Gephyrocapsa complex (Haptophyta). Protist 166, 323–336 (2015).
53. Gibbard, P. & Van Kolfschoten, T. in NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-019-12169-7 | www.nature.com/naturecommunications

54. Marlow, J., R., Lange, C. B., Wefer, G. & Rosell-Mele, A. Upwelling intensification as part of the Pliocene-Pleistocene climate transition. Science 290, 2288–2291 (2000).

55. Herbert, T. D., Peterson, L. C., Lawrence, K. T. & Liu, Z. H. Tropical ocean temperatures over the past 3.5 million years. Science 328, 1530–1534 (2010).

56. Diester-Haass, L., Billups, K. & Lear, C. Productivity changes across the mid-Pleistocene climate transition. Earth-Sci. Rev. 179, 372–391 (2018).

57. Lear, C. H. et al. Breathing more deeply: Deep ocean carbon storage during the mid-Pleistocene climate transition. Geology 44, 1035–1038 (2016).

58. Bard, E. & Rickaby, R. E. M. Migration of the subtropical front as a modulator of glacial climate. Nature 460, 380–383 (2009).

59. Galarza, J. A. et al. The influence of oceanographic fronts and early-life-history traits on connectivity among littoral fish species. Proc. Natl Acad. Sci. USA 106, 1473–1478 (2009).

60. Castelney, G. et al. Limits to gene flow in a cosmopolitan marine planktonic diatom. Proc. Natl Acad. Sci. USA 107, 12952–12957 (2010).

61. Aurahs, R., Grimm, G. W., Hemleben, V., Hemleben, C. & Kucera, M. Geographical distribution of cryptic genetic types in the planktonic foraminifer Globigerinoides ruber. Mol. Ecol. 18, 1692–1706 (2009).

62. Crampton, J. S. et al. Southern Ocean phytoplanктon turnover in response to stepwise Antarctic cooling over the past 15 million years. Proc. Natl Acad. Sci. USA 113, 6868–6873 (2016).

63. Vrba, E. Turnover-pulses, the Red Queen, and related topics. Am. J. Sci. 293, 418–452 (1993).

64. Henderiks, J. & Pagani, M. Refining ancient carbon dioxide estimates: Significance of coccolithophore cell size for alkaline-based PC02 records. Paleoceanography 22, 1–12 (2007).

65. Keller, M. D., Selvin, R. C., Claus, W. & Guillard, R. R. L. Media for the culture of oceanic ultraphytoplankton. J. Phycol. 23, 633–638 (2007).

66. Bollmann, J. Morphology and biogeography of Gephyrocapsa coccoliths in Holocene sediments. Mar. Micropaleontol. 29, 319–350 (1997).

67. Maniatis, T., Fritsch, E. F. & Sambrook, J. Molecular cloning: a laboratory manual. (Cold Spring Harbor Laboratory, 1982).

68. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).

69. Read, Ba et al. Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature 499, 209–213 (2013).

70. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).

71. McMcken, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).

72. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313 (2014).

73. Mirarab, S. et al. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30, i541–i548 (2014).

74. Carrier, G. et al. Draft genomes and phenotypic characterization of Tischysiris lutea strains. Toward the production of domesticated strains with high added value. Algal Res 29, 1–11 (2018).

75. Keeling, P. J. et al. The Marine Micrabort Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 12, e1001889 (2014).

76. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16, 157 (2015).

77. Kato, H., Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780 (2013).

78. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24, 1586–1591 (2007).

79. Dos Reis, M., Zhu, T. & Yang, Z. The impact of the rate prior on Bayesian estimation of divergence times with multiple loci. Syst. Biol. 63, 555–565 (2014).

80. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. 67, 901–904 (2018).

81. Bouckaert, R. R. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics 26, 1372–1373 (2010).

82. Schliep, K. P. phangorn: phylogenetic analysis in R. Bioinformatics 27, 592–593 (2011).

83. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290 (2004).

84. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704 (2003).

85. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J. R. Stat. Soc. B Met 57, 289–300 (1995).

86. Rannala, B. & Yang, Z. H. Baysian estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164, 1645–1656 (2003).

87. Okada, H. Neogene and Quaternary calcareous nanofossils from the Blake Ridge, Sites 994, 995, and 997. Proc. Ocean Drill. Program 164, 331–341 (2000).

88. Baumann, K. H. & Freitag, T. Pleistocene fluctuations in the northern Benguela Current system as revealed by coccolith assemblages. Mar. Micropaleontol. 52, 195–215 (2004).

89. Hagino, K. & Kulhanek, D. K. Data report: calcareous nanofossils from upper Pliocene and Pleistocene, Expedition 306 Sites U1313 and U1314. Proc. IODP 303/306, 1–5 (2009).

Acknowledgements
This work was supported by a grant from the John Fell Fund (Grant 152/079) to D.A.F. and R.E.M.R. and B.B.S.R.C. grant (BB/P009808/1) to D.A.F. The authors thank Prof Matsuoka (Kochi University, Japan) and Prof Baumann (University of Bremen, Germany) for providing Gephyrocapsa coccolith fossil data and the staff at the WTCGH (Oxford) for high throughput sequencing and initial data processing.

Author contributions
D.A.F. and R.E.M.R. conceived the project. E.M.B. and E.L.Y.W. prepared the samples for sequencing. E.M.B. and I.P. conducted SEM analyses. E.M.B., B.N. and D.A.F. conducted the analyses of sequence data. E.M.B., K.H. and J.R.Y. analysed and interpreted stratigraphic data. D.A.F. and E.M.B. wrote the paper with contributions from R.E.M.R. All authors contributed to editing the paper.

Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-12169-7.

Competing interests: The authors declare no competing interests.

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

Peer review information Nature Communications thanks Karl-Heinz Baumann, Iker Irigarri and Raphael Morard for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.