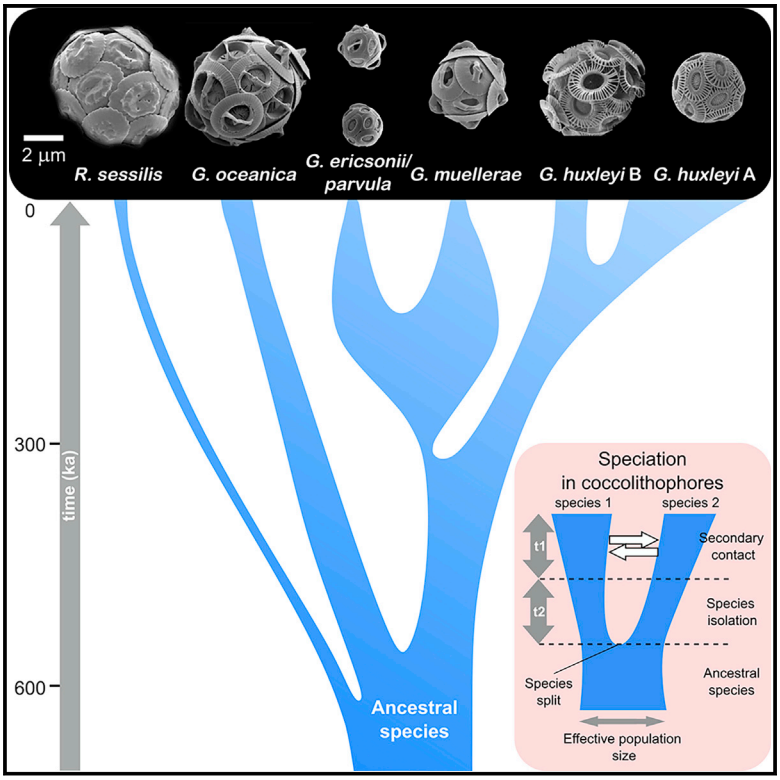
Article

Current Biology

The mode of speciation during a recent radiation in open-ocean phytoplankton

Graphical abstract



Highlights

- A combination of genetic and fossil data reveals speciation in coccolithophores
- Speciation tends to occur via complete isolation followed by secondary contact
- New species often form during the glacial periods of glacial-interglacial cycle
- Speciation in the open ocean is likely driven by extrinsic barriers to gene flow

Authors

Dmitry A. Filatov, El Mahdi Bendif, Odysseas A. Archontikis, Kyoko Hagino, Rosalind E.M. Rickaby

Correspondence

dmitry.filatov@plants.ox.ac.uk

In brief

Filatov et al. report the analysis of 43 coccolithophore genomes that reveals how new phytoplankton species form in the open ocean. Coccolithophore speciation tends to occur without gene flow, sometimes followed by secondary contact. It is likely driven by extrinsic physical barriers associated with glacial-interglacial cycles of the Pleistocene.

Filatov et al., 2021, Current Biology *31*, 1–11
December 20, 2021
https://doi.org/10.1016/j.cub.2021.09.073

CellPress

Please cite this article in press as: Filatov et al., The mode of speciation during a recent radiation in open-ocean phytoplankton, Current Biology (2021), https://doi.org/10.1016/j.cub.2021.09.073

Current Biology

CellPress
OPEN ACCESS

Article

The mode of speciation during a recent radiation in open-ocean phytoplankton

Dmitry A. Filatov,[1,4,*] El Mahdi Bendif,[1,2] Odysseas A. Archontikis,[2] Kyoko Hagino,[3] and Rosalind E.M. Rickaby[2]
[1]Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK
[2]Department of Earth Sciences, University of Oxford, South Parks Road, Oxford OX1 3AN, UK
[3]Centre for Advanced Marine Core Research, Kochi University, Nankoku, Kochi 783-8502, Japan
[4]Lead contact
*Correspondence: dmitry.filatov@plants.ox.ac.uk
https://doi.org/10.1016/j.cub.2021.09.073

SUMMARY

Despite the enormous ecological importance of marine phytoplankton, surprisingly little is known about how new phytoplankton species originate and evolve in the open ocean, in the absence of apparent geographic barriers that typically act as isolation mechanisms in speciation. To investigate the mechanism of open-ocean speciation, we combined fossil and climatic records from the late Quaternary with genome-wide evolutionary genetic analyses of speciation in the ubiquitous and abundant pelagic coccolithophore genus *Gephyrocapsa* (including *G. huxleyi*, formerly known as *Emiliania huxleyi*). Based on the analysis of 43 sequenced genomes, we report that the best-fitting scenario for all speciation events analyzed included an extended period of complete isolation followed by recent (Holocene) secondary contact, supporting the role of geographic or oceanographic barriers in population divergence and speciation. Consistent with this, fossil data reveal considerable diachroneity of species first occurrence. The timing of all speciation events coincided with glacial phases of glacial-interglacial cycles, suggesting that stronger isolation between the ocean basins and increased segregation of ecological niches during glaciations are important drivers of speciation in marine phytoplankton. The similarity across multiple speciation events implies the generality of this inferred speciation scenario for marine phytoplankton.

INTRODUCTION

How do new species emerge? This question is particularly acute in relatively homogeneous habitats, such as the open ocean, where few physical barriers to gene flow seem to exist[1] to promote allopatric speciation—divergence of physically isolated populations[2,3] was traditionally regarded as the predominant mode of speciation.[3] Speciation research remains heavily biased toward terrestrial organisms,[4] and the lack of data regarding the mode of speciation is particularly acute for pelagic protists (but see Postel et al.[5]). Despite their important and global contribution to the earth ecosystem, evolutionary genetic processes underpinning origination and adaptation in planktonic unicellular species remain poorly understood.[6] Constant mixing of pelagic plankton populations by ocean currents may lead to extensive gene flow across the species range. The homogenizing effect of this mixing may be difficult to overpower with diversifying selection, making physical barriers to gene flow essential for the establishment of new species.[7] We test this idea with genome-wide evolutionary genetic analysis in an ecologically important phytoplankton group—the coccolithophores.

Among the vast diversity of eukaryotic phytoplankton lineages, the coccolithophore family Noëlaerhabdaceae used in this study represents an ideal system for the study of speciation in marine phytoplankton. They are widely distributed, abundant, culturable, have relatively small genomes, and undergo intermittent sexual reproduction, with haploid and diploid generations in the life cycle.[8–10] The fossil record reveals that members of this family dominated coccolithophore populations over the last 20 million years (Ma),[11,12] with important implications for the global carbon cycle.[13,14] Coccolithophores produce more than 1 billion tonnes of $CaCO_3$/year,[15] which can accumulate as sediments and act as a major long-term geological sink of carbon dioxide. An integrated analysis of fossil and genome sequence data demonstrated that the extant members of this group, represented mostly by *Gephyrocapsa* species, originated in a species radiation within the last half a million years.[16] That study reported a macro-evolutionary pattern of cyclical coccolith size changes, with the radiation of extant species corresponding to the most recent of these cycles. The current study focuses on the micro-evolutionary processes during this species radiation.

Of the *Gephyrocapsa* species, the largest, *G. oceanica* (Figure 1), is usually found in mesotrophic subtropical and tropical waters. *G. muellerae*, which is medium sized, occurs in cooler productive waters. *G. ericsonii*, the smallest, is mostly reported in subtropical and tropical waters, occasionally in co-occurrence with the small ex-*Reticulofenestra* complex represented by *G. parvula*.[16,17] While most species used in this study formally belong to the *Gephyrocapsa* genus, we also included an outgroup *Reticulofenestra sessilis*, which, as shown below, is closely related to *Gephyrocapsa*. *R. sessilis* is a curious coccolithophore that is known to form symbiotic colonies around a diatom (Figure 1) in the deep-photic zone

Please cite this article in press as: Filatov et al., The mode of speciation during a recent radiation in open-ocean phytoplankton, Current Biology (2021), https://doi.org/10.1016/j.cub.2021.09.073

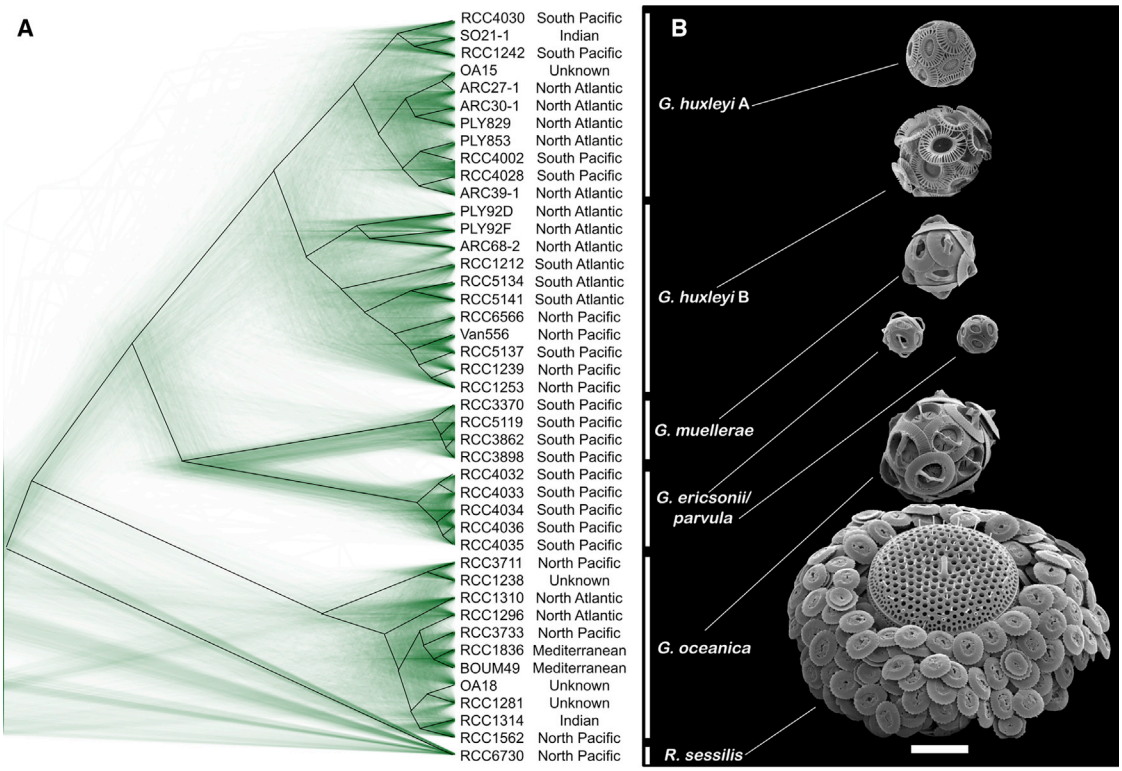


Figure 1. Phylogenetic relationships and gene-tree discordance of the Noëlaerhabdaceae

(A) DensiTree plot (green lines) based on 910 phylogenies constructed for genomic fragments 10 kb long and consensus species topology tree (black lines) inferred with a multi-coalescent method. All nodes are supported with >0.9 ASTRAL[26] branch value support (Figure S2).

(B) Scanning electron micrographs (SEM) of species used in the study. *R. sessilis* (strain RCC6730) are coccolithophores surrounding a symbiotic diatom in the center. The diatom is not analyzed in this study. SEMs are courtesy of Jeremy R. Young. For all SEMs, the scale bar represents 4 μm. See also Figures S1–S3 and Table S1.

of equatorial and tropical oceans.[18,19] This species is little studied and has not been isolated previously. It is worth noting that our analysis also includes *Gephyrocapsa huxleyi*, formerly known as *Emiliania huxleyi*. Consistent with this name change, this species clusters within *Gephyrocapsa*, as shown below. *G. huxleyi* has become a model species for marine phytoplankton studies, and many strains are available in culture collections.[20–22] Its ~140-megabase (Mb) genome is sequenced,[23] and basic population genetic parameters have already been characterized to some extent.[24] This emblematic coccolithophore species is ubiquitous and so abundant in the modern oceans that the coccoliths shed during the extensive annual blooms are visible from space. It is thought to be one of the main calcite producers on Earth[21] and plays an important role in the global carbon cycle.[25]

In this study, we combined population genetic analysis of whole-genome sequence data from 43 *Gephyrocapsa* and *Reticulofenestra* isolates from across the global ocean, with fossil and climatic records from the late Quaternary to shed light on the evolutionary processes underpinning speciation in marine phytoplankton.

RESULTS

The strains and sequence data used in the analyses

Our analyses are based on 28 newly sequenced and 15 previously published genomes for *Gephyrocapsa* and *Reticulofenestra* clonal isolates from all over the world oceans (Figure 1A; Table S1). All the strains in the analysis were sequenced with short read Illumina paired end sequencing, totaling 360.9 gigabase (Gb) of newly generated sequence data across all the strains (Table S1). The ploidy of strains was checked with K-mer spectra,[27,28] which consistently formed two peaks, indicating diploidy (Figure S1).

The analyzed set of strains includes two morphotypes (A and B) described for *G. huxleyi*,[29] as well as nearly all known extant morphospecies in the genus *Gephyrocapsa* (Figure 1B). Due to considerable genetic differentiation between the strains belonging to the A and B morphotypes of *G. huxleyi* (Figure 1; Table 1), these morphotypes are analyzed separately and referred to as *G. huxleyi* A and B, respectively.

Sequence diversity within and between coccolithophore species

The total sequence diversity in our sample of *Gephyrocapsa* and *Reticulofenestra* strains (pooling all strains of all species together), calculated as the average number of nucleotide differences per synonymous site ($\pi = 0.039 \pm 0.0159$), reveals that analyzed species are closely related. The total genetic diversity at non-synonymous sites is over three times lower than those per synonymous site ($\pi = 0.013 \pm 0.0057$), indicating the presence of considerable purifying selection. Analysis of molecular variance (AMOVA)[30] applied to species with multiple individuals

Please cite this article in press as: Filatov et al., The mode of speciation during a recent radiation in open-ocean phytoplankton, Current Biology (2021), https://doi.org/10.1016/j.cub.2021.09.073

Table 1. Sequence divergence and population differentiation for all pairs of species

Species	A	B	Gm	Gp	Ge	Go	Rs
A	0.0085[a]	0.0153[b]	0.0302[b]	0.0334[b]	0.0251[b]	0.0608[b]	0.0243[b]
B	0.3602[c]	0.0085[a]	0.0293[b]	0.0317[b]	0.0238[b]	0.0597[b]	0.0236[b]
Gm	0.7556[c]	0.7490[c]	0.0018[a]	0.0268[b]	0.0201[b]	0.0548[b]	0.0215[b]
Gp	0.7692[c]	0.7573[c]	0.8023[c]	0.0019[a]	0.0011[b]	0.0579[b]	0.0238[b]
Ge	0.5437[c]	0.5193[c]	0.5501[c]	N/A[c]	0.0019[a]	0.0434[b]	0.0179[b]
Go	0.8378[c]	0.8354[c]	0.8642[c]	0.8658[c]	0.7349[c]	0.0061[a]	0.0242[b]
Rs	0.4563[c]	0.4431[c]	0.4991[c]	0.5342[c]	N/A[c]	0.4533[c]	0.0059[a]

Intra-specific diversity for species with just a single sample sequenced (Ge and Rs) are based on heterozygosity of that sample. See also Table S2. A and B, *G. huxleyi* morphotypes A and B, respectively; Ge, *G. ericsonii*; Gm, *G. muellerae*; Go, *G. oceanica*; Gp, *G. parvula*; N/A, not available; Rs, *R. sessilis*.

[a]Synonymous genetic diversity within species (π)

[b]Synonymous (4-fold degenerate) pairwise sequence divergence between the species (D_{xy})

[c]Population differentiation (F_{st}; all significant by permutation test, $p < 0.001$)

sequenced reveals that about 88.6% of total diversity is due to divergence between the species and only about 11.4% due to polymorphism within the species (Table S2).

Although most of the genetic diversity in this group is due to species divergence (Table S2), the species analyzed are closely related and synonymous nucleotide divergence between the most divergent species *G. oceanica* and most other species is only about 6% (Table 1), although divergence between the other species analyzed is even lower. Population (species) differentiation, measured with F_{st}, is high and significant for all comparisons of species with more than one sample sequenced (Table 1, below diagonal). This conclusion is also visually supported by the DensiTree plot showing clustering of the samples by species at different genes across the genome (Figure 1A).

The analyses within the species reveal relatively low nucleotide polymorphism (Table 1, diagonal), consistent with the previous report of low diversity in *G. huxleyi*.[24] Genetic diversity in globally distributed *G. huxleyi* (both A and B morphotypes separately) and *G. oceanica* is similar, while the diversity in *G. muellerae*, *G. parvula*, and *G. ericsonii* is much lower (Tables 1 [diagonal] and 2), consistent with their limited distribution. It is interesting that genetic diversity (heterozygosity) in the single *R. sessilis* isolate sequenced (π = 0.59% ± 0.587%) is closer to the level of polymorphism in the ubiquitous *G. huxleyi* and *G. oceanica* rather than the geographically restricted *G. muellerae*, *G. parvula*, and *G. ericsonii*. It is worth noting that smaller sample sizes for these species do not affect the accuracy of their genetic diversity estimates because different regions across the genome provide independent samples of the past coalescent process; as such, it is more informative to analyze longer sequences than larger sample sizes.[31,32]

The models of speciation with isolation and secondary contact fit the data best

In order to analyze the population genetic processes during consecutive speciation events in this ecologically important phytoplankton group, we used a wide range of population split speciation models, allowing for population size change and interspecific gene flow (Figure 2A; Table S3), implemented in a Poisson random-field framework.[35] The models were fitted to genome-wide polymorphism data from pairs of species with at least four diploid individuals sequenced. The model fit was quantified with log-likelihood, and the models were compared using sample-size-corrected Akaike information criterion (AICc).[36] Comparing the model fit for simpler and more complex models, we tested whether additional parameters significantly improve the model fit to data (Table S4). Overall, we analyzed model fit and estimated parameters for 30 different speciation models in pairwise species analyses (Figure 2A; Table S4). This revealed that secondary contact (SC) models, allowing for a period of isolation followed by secondary contact (Figure 2B), fitted the data best (Figure 2A; Table S4). The parameter estimates for the best-fitting model for each speciation event analyzed are listed in Table 3, and the inferred speciation scenario is summarized in Figure 2D. To convert the parameter estimates to biologically meaningful units, we used the per-nucleotide mutation rate ($\mu = 5.5 \times 10^{-10}$ [SD: 5.05×10^{-10}–6.09×10^{-10}] per nucleotide per cell division) measured for *G. huxleyi* in a recent mutation accumulation experiment.[37]

The oldest speciation event analyzed (labeled "1" in Figure 2D) includes the split between *R. sessilis* and all other species in this clade. *R. sessilis* is a previously unsequenced species that evolved a symbiotic relationship with a centric diatom (Figure 1B), which might have caused the speciation of *R. sessilis*. According to pairwise sequence divergence, *R. sessilis* is equidistant from other species analyzed (Table 1), indicating its basal position in this clade. The chloroplast genome sequences from outgroups *Isochrysis galbana* and *Tisochrysis lutea* confirm the basal position of *R. sessilis* among the Noëlaerhabdaceae species (Figure S3). The DensiTree plot (Figure 1A) reveals that some *R. sessilis* genes cluster with *G. oceanica*, while other genes are closer to the clade including all other species analyzed. This may reflect incomplete lineage sorting or some introgression between these species. Despite the basal position of *R. sessilis*, its divergence from other species ($d_{xy} \sim 0.025$) is two times lower than that for *G. oceanica* ($d_{xy} \sim 0.055$; Table 1), which suggests slower divergence of *R. sessilis*, likely due to a longer generation time compared to the other species studied here. Indeed, *R. sessilis* inhabits the deep photic zone in the ocean and is characterized by a relatively slow growth rate in culture. As only one *R. sessilis* sample is available, we cannot infer the parameters of speciation using the polymorphism-based

Please cite this article in press as: Filatov et al., The mode of speciation during a recent radiation in open-ocean phytoplankton, Current Biology (2021), https://doi.org/10.1016/j.cub.2021.09.073

Table 2. Level and patterns of genetic diversity at different types of sites

Species	n[a]	L_s[b]	L_n[c]	S_s[d]	S_n[e]	p_s (%)[f]	p_n (%)[g]	D_{Taj}[h]	Z_{nS}[i]
A	11	606,144	2,532,365	20,100	34,890	0.85 ± 0.450	0.33 ± 0.187	−0.267	0.096
B	11	606,144	2,532,365	21,575	38,557	0.85 ± 0.483	0.34 ± 0.207	−0.520	0.087
Gm	4	606,144	2,532,365	2,068	5,036	0.18 ± 0.072	0.11 ± 0.042	2.116[j]	0.416
Gp	4	606,144	2,532,365	2,544	6,242	0.19 ± 0.088	0.11 ± 0.052	0.876	0.260
Ge	1	606,144	2,532,365	1,163	2,572	0.19 ± 0.192	0.10 ± 0.102	N/A	N/A
Go	11	606,144	2,532,365	16,665	25,183	0.61 ± 0.373	0.25 ± 0.135	−0.795	0.109
Rs	1	606,144	2,532,365	3,559	9,545	0.59 ± 0.587	0.38 ± 0.377	N/A	N/A
Total	43	606,144	2,532,365	101,342	152,025	3.88 ± 1.587	1.28 ± 0.570		

Species abbreviations are as in Table 1.
[a]Sample size
[b]Synonymous positions analyzed
[c]Non-synonymous positions analyzed
[d]Synonymous polymorphic sites
[e]Non-synonymous polymorphic sites
[f]Average heterozygosity per 100 synonymous sites
[g]Average heterozygosity per 100 non-synonymous sites
[h]Tajima's D[33] at synonymous sites. The positive D_{Taj} values for *G. muellerae* and *G. parvula* likely reflect population contraction of these species (Figure 2D).
[i]Z_{nS}[34] at all sites calculated as weighted average for 260 longest genomic contigs, with weighting proportional to the length of the contig.
[j]$p < 0.05$

approach, which we have employed for other species splits. Nevertheless, given that heterozygosity in the sequenced *R. sessilis* sample is similar to that in *G. oceanica* (Table 2), the population size of this species is likely comparable to that in *G. oceanica*. The sediment data (Figures 2E and 2F) also show comparable species abundance for *G. oceanica* and *R. sessilis*, though *G. oceanica* abundance shows occasional peaks (e.g., at around 220 ka), while the abundance of *R. sessilis* is more stable through time.

The second speciation event (labeled "2" in Figure 2D) was analyzed using the two species pairs with the largest sample sizes available: *G. huxleyi* A–*G. oceanica* (A-Go) and *G. huxleyi* B–*G. oceanica* (B-Go) (Figure 2A; Tables 3 and S4). Based on the best-fitting SC model, the speciation is estimated to have occurred around 565 ka (Table 3). We estimate that *G. huxleyi* and *G. oceanica* lineages inherited more or less an even proportion of ancestral species diversity ($0.3 < s < 0.6$ in A-Go and in B-Go analyses, respectively; Table 3). Following the species split, the population size of *G. oceanica* remained stable over time, while the population of *G. huxleyi* has grown considerably in recent times (Figure 2D), an inference that is consistent with the sediment record of increasing *G. huxleyi* abundance toward the modern day (Figures 2E and 2F). Although abundance of a species in sediment does not directly reflect its past population size, the increased abundance at multiple sediment locations likely reflects expanding population size.

The third speciation event (labeled "3" in Figure 2D; ~300 ka, Table 3) gave rise to the clade including *G. huxleyi* species A and B (hereafter AB) and the "Gmpe"—the lineage including *G. muellerae* ("Gm"), *G. parvula* ("Gp"), and *G. ericsonii* ("Ge"). Our analysis of this species split involved four species pairs: A-Gm; B-Gm; A-Gp; and B-Gp (Figure 1A; Tables 3 and S4). While the small sample sizes for Gm and Gp could affect the accuracy of the analysis, the consistency of the parameter estimates across the species pairs (Table 3) provides some reassurance in the accuracy of the analyses for this speciation event. According to the analyses including either *G. huxleyi* A or B and either Gm or Gp, the speciation of AB and Gmpe occurred with over 95% of the ancestral diversity inherited by the AB lineage ($s > 0.95$; Table 3) and <5% going into the Gmpe lineage (Figure 2D). This species split was followed by population growth in *G. huxleyi* and population growth and then decline for both *G. parvula* and *G. muellerae* to their current sizes, an order of magnitude smaller than the ancestral species at the Gmpe-AB species split (Figure 2D), consistent with the current low genetic diversity in *G. parvula* and *G. muellerae* (Table 2).

The scenario of population expansion, followed by a decline in the Gmpe lineage, is also supported by the parameter estimates in the IMpre model that allows for a population size change prior to speciation. Applying that model to the species split between *G. parvula* and *G. muellerae* (labeled "4" in Figure 2D), we estimated that the effective population size in the Gmpe lineage expanded from the ancestral ~4.8 million to 34.5 million before speciation and then underwent a massive decline after the species split 4. A similar scenario can be inferred using the best-fitting SCm1_hn model, which estimates that the ancestral effective population size was ~16.8 million just before the species split 4 that occurred around 167.1 ka, followed by a massive decline in both *G. parvula* and *G. muellerae* (Table 3). The sediment data also support this scenario, with a peak in *G. muellerae* abundance around 150 ka (Figure 2E). After the first common occurrence of *G. muellerae* in the fossil record at marine isotope stages (MIS) 6/7 (c. 190–170 ka),[42,43] this species dominated the coccolith populations for about 50 ka, but its abundance declined as *G. huxleyi* abundance grew, exceeding the *G. muellerae* population by the MIS 4/5 boundary (~71 ka).[44–47] Our estimated time of *G. parvula* and *G. muellerae* speciation is more recent than the *G. muellerae* occurrence in

Please cite this article in press as: Filatov et al., The mode of speciation during a recent radiation in open-ocean phytoplankton, Current Biology (2021), https://doi.org/10.1016/j.cub.2021.09.073

CellPress
OPEN ACCESS

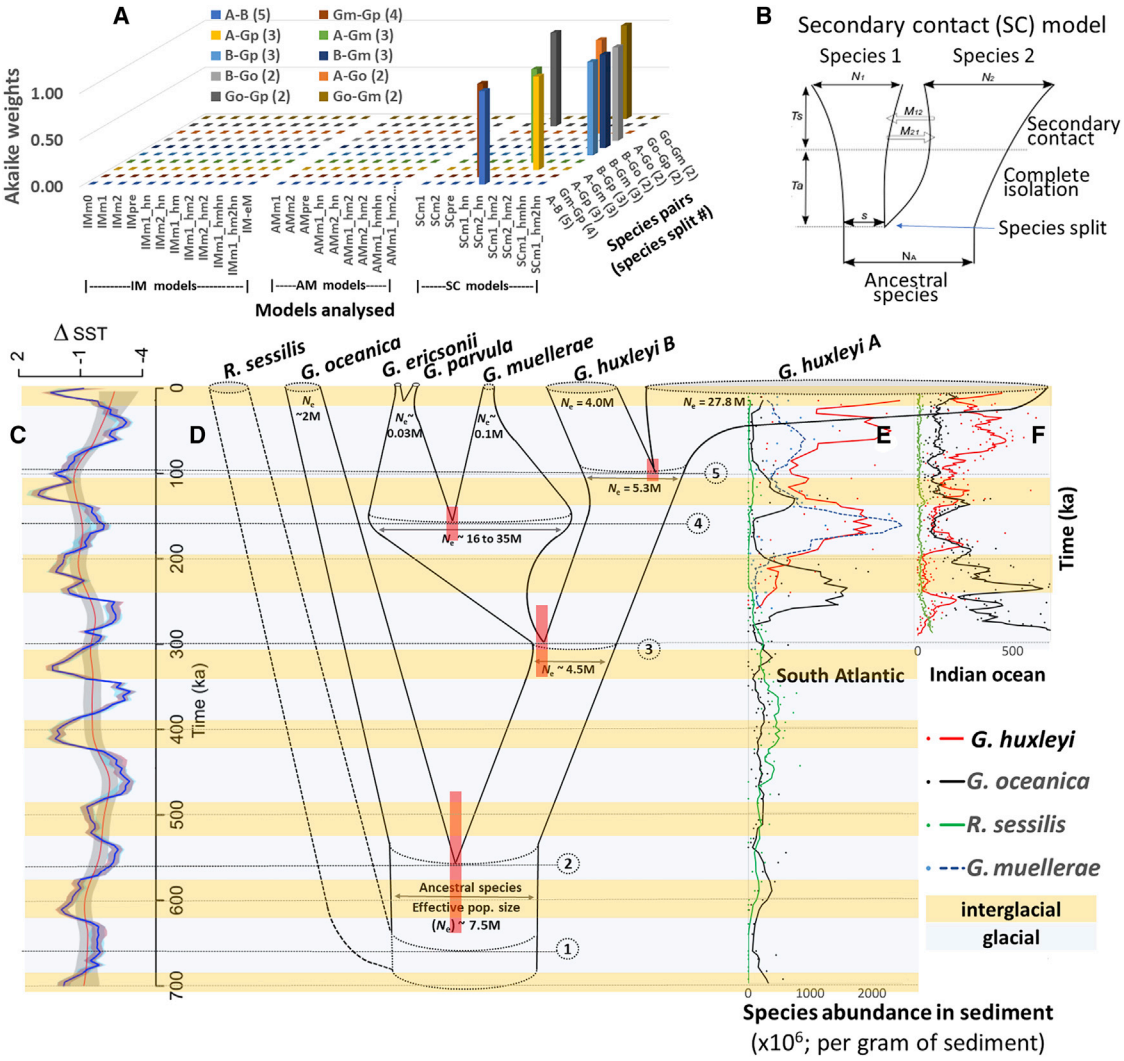


Figure 2. Reconstruction of demographic history of speciation in coccolithophore family Noëlaerhabdaceae

(A) Akaike weights (formula 4 in Wagenmakers and Farrell[38]) for 30 speciation models fitted to ten species pairs (representing four species splits) show that secondary contact models (SC…) fit the data best. The details of different models are described in the STAR Methods and Table S3.

(B) The diagram of secondary contact model. The ancestral species with effective size N_a splits into two derived species, with the proportion s of ancestral diversity inherited by the species 1 and 1-s inherited by the species 2, and over time, population sizes of the two species change exponentially to reach the present sizes N_1 and N_2. Following the species split, the two species remain completely isolated for the time period T_1, but interspecific gene flow restarts after secondary contact at time T_2 in the past. Interspecific migration after secondary contact occurs with rates (M_1 and M_2) in two directions.

(C) The change of sea surface temperature (SST) through time.[39]

(D) Visual representation of the parameters inferred for consecutive speciation events in Noëlaerhabdaceae. The width of the branches reflects effective population size (N_e) estimated for the extant and ancestral species. The red bars at the species splits show min-max ranges for the estimates of the speciation time across bootstrap replicates, taking mutation rate uncertainty (from Krasovec et al.[37]) into account. The actual parameter estimates are listed in Table 3.

(E) The counts of coccoliths (×10^6; per gram of sediment analyzed) from Ocean Drilling Program (ODP) site 1082 from the eastern South Atlantic off Namibia.[40] The counts for *G. muellerae* are scaled down 5-fold to fit the high peak around 150 ka to the scale of the plot. The colored dots are the actual data, and the lines show moving average of five data points for *G. huxleyi* (red), *G. oceanica* (black), *R. sessilis* (green), and *G. muellerae* (dashed blue).

(F) The counts of coccoliths (×10^6; per gram of sediment analyzed) from core GeoB12613-1 from the western Indian Ocean off Tanzania.[41]

Yellow and blue shading across (C)–(F) correspond to interglacial and glacial periods, respectively. See also Figures S4 and S5 and Tables S3 and S4.

the fossil record, which may be explained by the evolution of a phenotype resembling *G. muellerae* in the ancestral species before the species split between *G. parvula* and *G. muellerae*. After that speciation event, *G. muellerae* retained the phenotype of the ancestral species, while *G. parvula* lost the coccolith bridge and evolved its modern phenotype (Figure 1B).

The split between *G. huxleyi* species A and B (labeled "5" in Figure 2C) is estimated to have occurred around 99.4 ka (Table 3), with roughly a quarter of the ancestral diversity inherited by *G. huxleyi* A ($s \sim 0.2$; Table 3) and the rest by *G. huxleyi* B. The latter is inferred to have had a stable population size since speciation, while the former has grown considerably (Figure 2D; Table 3).

Please cite this article in press as: Filatov et al., The mode of speciation during a recent radiation in open-ocean phytoplankton, Current Biology (2021), https://doi.org/10.1016/j.cub.2021.09.073

Table 3. Parameters of the best fit speciation models

Species[a]	s	N_a	N_1	N_2	T_{spec} (ka)	T_{SC} (ka)	M	Best fit model
A-B	0.2 (0.17–0.23)[b]	5.3 (4.78–5.99)	27.8 (19.23–50.23)	3.9 (3.16–4.49)	99.4 (86.25–108.55)	9.9 (6.74–10.99)	0.35 (0.35–0.42)	SCm2_hn
A-Go	0.5 (0.48–0.51)	7.7 (9.74–12.39)	9.3 (8.59–9.97)	2.1 (2.06–2.59)	572.6 (519.4–625.8)	12.8 (10.6–14.4)	0.14 (0.23–0.27)	SCm2_hm2
B-Go	0.6 (0.54–0.71)	7.2 (6.08–8.25)	3.98 (0–32.6)	1.29 (0.49–1.59)	565.1 (474.5–655.8)	9.45 (5.17–22.57)	0.16 (0.05–0.16)	SCm1_hm2hn
A-Gp	0.95 (0.92–0.97)	6.7 (5.79–7.58)	5.9 (4.92–6.92)	6.1 (1.41–23.04)	316.4 (270.2–352.4)	10.5 (3.86–17.55)	0.19 (0.27–0.61)	SCm2_hm2
A-Gm	0.98 (0.97–0.99)	4.9 (3.46–5.48)	6.8 (5.84–7.83)	4.4 (2.38–38.66)	296.8 (269.6–357.0)	1.9 (1.45–4.24)	0.62 (0.32–0.74)	SCm1_hm2
B-Gp	0.98 (0.94–0.98)	2.9 (2.14–6.29)	0.9 (0.47–5.45)	2.2 (0.58–5.81)	368.3 (288.6–417.6)	15.2 (2.87–19.24)	0.11 (0.04–0.34)	SCm1_hm2hn
B-Gm	0.97 (0.96–0.99)	3.7 (1.59–5.85)	1.9 (0.58–9.98)	0.9 (0.26–5.65)	309.8 (249.9–368.9)	6.3 (0.96–14.85)	0.39 (0.04–0.90)	SCm1_hm2hn
Gm-Gp	0.5 (0.34–0.75)	16.2 (13.9–17.6)	0.03 (0.014–0.206)	0.05 (0.014–0.273)	167.1 (156.9–177.2)	13.7 (11.88–23.13)	0.004 (0.00–0.02)	SCm1_hn

See also Tables S3 and S4. M, interspecific migration from sp2 to sp1 and from sp1 to sp2 averaged across the genome and the direction of migration; N_a, N_1, N_2, estimated effective population size ($*10^6$) of the ancestral species and extant species 1 and 2, respectively; s, the proportion of ancestral species inherited by sp1; 1-s is inherited by sp2; T_{spec}, T_{sc}, time of speciation and secondary contact, respectively (thousands of years ago, ka).

[a]Species abbreviations are as in Table 1

[b]Confidence intervals (in parentheses) show the minimal and maximal estimates for 100 bootstrap replicates, taking the uncertainty of mutation rate estimate ($m = 5.05–6.09 * 10^{-10}$ per nucleotide per cell division)[37] into account. That is, the minimal and maximal parameter estimates were calculated assuming the mutation rates $m = 6.09 * 10^{-10}$ and $5.05 * 10^{-10}$, respectively.

This is consistent with the *G. huxleyi* sediment record (Figures 2E and 2F) that indicates a massive increase in population size in the last 100 ka.[42,44,46,48] Following the appearance of *G. huxleyi* in the fossil record ~290 ka,[12] the abundance of *G. huxleyi* has grown dramatically to current estimated census size of 7×10^{22} living cells[49] but fluctuated over time (Figures 2E and 2F). The effective size of a fluctuating population is equal to a harmonic mean of population sizes over time,[50,51] which explains why the expansion of the *G. huxleyi* effective population size inferred in our population genetic analysis (Figure 2A) is smaller than implied by the rise in coccolith abundance in the sediment record. Furthermore, the 15 orders of magnitude disparity between this astronomical census size and estimated effective population size ($N_e \sim 10^7$; Table 3) is likely due to the effect of linked selection that limits genetic diversity in very large populations—an extreme instance of "Lewontin's paradox" analyzed previously in *G. huxleyi*.[24]

Hybridization between *Gephyrocapsa* species is likely a recent phenomenon

The ABBA-BABA-like tests for interspecific gene flow[52,53] detected significant gene exchange between most *Gephyrocapsa* species analyzed (Figure S4; Table S5). Consistent with this, the model without gene flow (IMm0) fitted data much more poorly compared to models allowing for gene flow (Table S4), indicating that interspecific gene flow is an important feature of the data. However, in all cases, gene flow was estimated to be low ($M < 1$; Table 3), revealing that interspecific hybridization between the species in this genus is not too common.

To analyze the evolution of barriers for interspecific gene flow, we used the models allowing for migration to change over time. In one such model, IM-eM, migration changes exponentially from migration rate right after speciation (M_{split}) to migration at present ($M_{present}$). This model consistently estimated the $M_{split} = 0$ and $0 < M_{present} < 0.1$. The other models—ancestral migration (AM) and SC—allowed for two time periods with distinct migration rates over time (Figure 2B). The AM models correspond to a scenario of species evolving reproductive isolation over time, with interspecific gene flow initially present after the species split but prevented by evolution of reproductive barriers later on. The SC models implement a secondary contact scenario when two previously isolated (sub)species meet and start to hybridize (Figure 2B). The SC models fitted the data significantly better than the alternative models (Figure 2A; Table S4), indicating that a period of complete isolation, followed by secondary contact and hybridization, is a plausible scenario for species in this phytoplankton group. The time of secondary contact, as estimated in the best-fitting SC models ($T_{sc} < 15$ ka in most species comparisons; Table 3), indicates that interspecific hybridization and gene exchange is a recent phenomenon in this group, possibly driven by climatic changes, such as the ending of the last ice age. The models allowing for multiple phases with variable migration (VM models) do not provide support for interspecific migration during older glacial cycles, possibly due to lack of power and convergence problems with too-parameter-rich models (Figure S5).

The secondary contact scenario implies that the species are initially isolated by extrinsic factors, such as geographic or oceanographic barriers, while maintaining some degree of

Please cite this article in press as: Filatov et al., The mode of speciation during a recent radiation in open-ocean phytoplankton, Current Biology (2021), https://doi.org/10.1016/j.cub.2021.09.073

CellPress
OPEN ACCESS

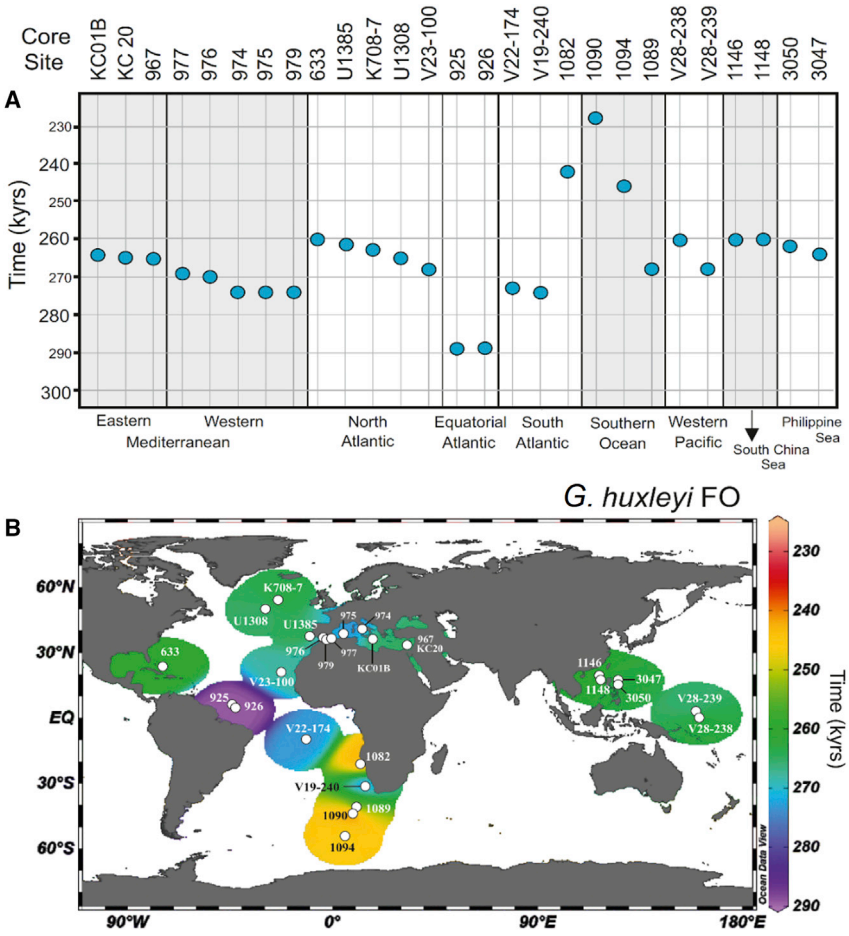


Figure 3. The diachroneity in first occurrence (FO) of *G. huxleyi* in sediment data across the world oceans

The timing of FO (thousands of years ago, kyrs) is shown on the y axis in (A) and with color in (B). White dots on the map (B) indicate the locations of the sediment cores, and the color reflects the time of the FO in the area around each sediment core. To ensure reliability of the pattern observed, we included only the areas with at least two cores analyzed in the area. The data sources are listed in Table S6. See also Table S7.

reproductive compatibility enabling interspecific gene flow after secondary contact. To shed light on the extrinsic factors that may have caused isolation and drove speciation in this phytoplankton group, we used fossil data available for this calcifying phytoplankton group.

Fossil-based evidence of speciation in Noëlaerhabdaceae

The origin of *G. huxleyi* is relatively well documented in the fossil record (Table S6). In order to analyze where the species first originated and whether geographic (e.g., continents or ice sheets) and physical oceanographic barriers (e.g., ocean fronts) acted as isolating mechanisms promoting speciation, we compiled the timing of the first occurrence (FO) of *G. huxleyi* across the global oceans, focusing on the regions with multiple sediment cores available to ensure the accuracy of FO dating (Figure 3). Diachroneity of FO in different places could support the role of physical isolation mechanisms helping genetic isolation and speciation. The data reveal that *G. huxleyi* first appeared in the tropics, and there is an apparent delay of 20 ka for *G. huxleyi* to escape from its origination at the equator into temperate latitudes and up to ~50 ka delay before the *G. huxleyi* FO reaches the polar regions (Figure 3). There is also an apparent delay of ~20–30 ka with *G. huxleyi* reaching the tropical regions of the Western Pacific, though the areas outside the Atlantic are too sparsely covered by core data to infer the patterns of species spread between the basins. Similar analyses of sediment data for other species in this clade are significantly more complicated due to difficulties in the exact placement of their biohorizons in the fossil record[54,55] and the different taxonomic criteria followed by authors (Table S7) for their identification.

DISCUSSION

Here, we reported population genetic analyses of speciation in a predominant marine phytoplankton group—the coccolithophores. Our previous analysis of this group revealed that extant diversity in *Gephyrocapsa* originated in a single radiation event that started only half a million years ago.[16] It also demonstrated that cycles in predominance of large and small coccolithophores in the fossil record correspond to consecutive events of species radiations followed by extinctions, with the extant diversity in this genus representing the latest species radiation. However, the previous work was based on phylogenetic analyses of only ten sequenced strains,[16] which was uninformative about the predominant speciation mode in this radiation. The analyses of 43 strains presented above allowed us to study the evolutionary genetic processes underpinning speciation in this ubiquitous and abundant phytoplankton group. The analyses revealed patterns in common for all speciation events, allowing us to reconstruct a speciation scenario for this phytoplankton group.

First, our results suggest that the formation of new species in marginal isolated populations is not the primary mechanism of speciation in oceanic phytoplankton. Given the lack of obvious physical barriers in the open ocean, origination of new plankton species in small marginal populations cut off from the sea seems a plausible possibility. This scenario implies strongly uneven population sizes at the species splits. Contrary to this, three out of four speciation events analyzed (numbered 2, 4, and 5 on Figure 2D) show relatively even species splits ($0.2 < s < 0.8$; Table 3), which is incompatible with speciation in marginal isolated populations.

Second, for all speciation events analyzed, the secondary contact models had by far the highest likelihood (Figure 2A).

Please cite this article in press as: Filatov et al., The mode of speciation during a recent radiation in open-ocean phytoplankton, Current Biology (2021), https://doi.org/10.1016/j.cub.2021.09.073

Our results are consistent with nearly instantaneous shutdown of gene flow at the time of speciation and an extended period of complete isolation, followed by recent secondary contact and restart of gene flow (Table 3). In the terrestrial realm, such speciation dynamics are often associated with vicariance driven by glacial cycles, with fragmentation of species ranges in glacial or interglacial refugia (e.g., on separate islands isolated by rising sea level). For example, in Europe, many closely related species formed during the last ice age due to isolation and divergence between refugia at the Iberian peninsula, Italy, Greece, and Anatolia.[56] In the marine realm, similar dynamics can be caused by cycles of opening and closing of straits by glaciers or changing of sea level.[57] The movement of oceanic fronts closer to the equator during the glacial periods[58] and poleward shifts in the interglacials may also act as one possible mechanism isolating the oceanic basins and driving the cycles of population isolation and mergers for marine phytoplankton.

How physical barriers to gene flow play a role in plankton speciation remains unclear, but the availability of an extensive *Gephyrocapsa* sediment record[40,41,59–61] (Table S6) provides an additional resource about the past history of this group of species. The fossil evidence points to the emergence of *G. huxleyi* at equatorial latitudes, suggesting that the genetic isolation that led to speciation was a low-latitude phenomenon, in accord with the suggestion that the tropics are a diversity pump.[62,63] The delay in *G. huxleyi* appearance at temperate and polar waters may reflect slow dispersal of the species after origination at equatorial latitudes. However, near synchronous appearance of *G. huxleyi* in the tropical and temperate regions of the Atlantic and the Western Pacific (substantially earlier than in the Southern Ocean; Figure 3) suggests that dispersal is unlikely to be the limiting factor for the spread of this species. Thus, it appears more likely that diachroneity in *G. huxleyi* FO could be caused by physical and/or ecological barriers. More detailed analysis of the sediment data from multiple regions is required to reconstruct past biogeography and the history of speciation and compare it to the results of evolutionary genetic analyses.

Third, all the analyzed speciation events occurred during the onset of glacial conditions at different times in the last ~0.6 Ma (Figure 2C), suggestive that the glacial ocean is more conducive to genetic isolation and the emergence of new species. However, given the limited number of speciation events analyzed, the coincidence of speciation with glacial conditions cannot be tested statistically and has to be taken with caution. In addition to the above-mentioned vicariance caused by glacial circles, the changes in ocean circulation and water column stratification may promote isolation and speciation in the glacial ocean. The growth of ice sheets during glacial periods, aggravated polar ocean cooling, and steepened latitudinal temperature gradients led to ventilation of the deep ocean by colder, denser waters sitting beneath the more muted cooling in the lower latitude regions. Conceptually, the segregation of ecological niches may have been more distinct within the glacial ocean. It experienced greater density contrasts both vertically and from low to high latitude: the lower latitude water column was more strongly stratified and stronger oceanic fronts were more compressed equatorward (e.g., Bard and Rickaby[58]). It remains to be studied whether depth stratification or ocean fronts can cause complete isolation of plankton species implied by the extended isolation phase in our best-fit SC models (Table 3).

Fourth, all species pairs showed a low but significant level of interspecific gene flow, consistent with the view that gene flow between nascent species is common in the marine realm.[64–66] However, the rate of gene flow was estimated to be fairly low ($M << 1$), suggesting that hybridization played minimal, if any, role in the evolution of these species. Complete reproductive isolation may take millions of generations to evolve after the initial species split. Gradual evolution of genetic incompatibilities between the species is expected to result in more pronounced heterogeneity of gene flow across the genome for older speciation events due to reduced gene flow in regions adjacent to genes causing species incompatibility. Indeed, the models allowing for heterogeneous migration (…_hm, …_hmhn and …_hm2hn) across the genome were the best-fit models for the older species splits 2 and 3, but not for the more recent speciation events 4 and 5 (Figure 2A; Table S4).

Fifth, for all speciation events analyzed, the parameter estimates for the best-fitting SC models (Table 3) indicate that the gene flow was restricted to a short period of secondary contact in the Holocene after an extended period of complete isolation. This result is consistent with the glacial cycle driving vicariance and secondary contact in the marine plankton, not dissimilar to that described for terrestrial organisms.[56] We hypothesize that the secondary contact and low-level hybridization between the coccolithophore species analyzed may have been induced by perturbation of marine ecosystems by the discharge of large amounts of fresh water from rapidly melting glaciers during the abrupt ending of the last ice age ~14 ka. Melting glaciers are known to fertilize the ocean with iron[67] and phosphorus,[68] which stimulates extensive phytoplankton blooms.[69] Such blooms left traces in barium content in sediments that peak at glacial to interglacial transitions, reflecting past deglacial productivity pulses.[70] Interspecific hybridization could have occurred during such phytoplankton blooms during glacial to interglacial transitions. Additionally, a more stratified glacial ocean may have stored more nutrients at depth, which become remixed throughout the water column on breakdown of stratification at the glacial termination.[70] Alternatively, the poleward migration of basin-isolating fronts (e.g., the subtropical front) beyond the tips of southerly continents on glacial terminations could also enhance genetic exchange among the coccolithophore population. Because each of these mechanisms is likely to occur on every deglacial transition, it may be hypothesized that glacial to interglacial transitions in the older glacial cycles have also driven bouts of interspecific hybridization. Detecting such events would require more complex models with multiple migration phases. However, such VM models show a poorer fit to data compared to the best-fit SC models, which may be due to lack of power to detect multiple migration events and problems with model convergence for too-parameter-rich models (Figure S5).

While we detected interspecific gene flow, it was limited to a short period of time following Holocene secondary contact; the speciation events (species splits) appear to have occurred without ongoing gene flow. The scenario of speciation without interspecific gene flow favored by our analyses appears to contradict the view that speciation with gene flow is a common

Please cite this article in press as: Filatov et al., The mode of speciation during a recent radiation in open-ocean phytoplankton, Current Biology (2021), https://doi.org/10.1016/j.cub.2021.09.073

mode of speciation in the marine realm.[4,5,65,66,71,72] However, this view is primarily based on studies in coastal, benthic marine organisms and the pelagic realm remains understudied. Furthermore, a careful look into population genetics of speciation may help to explain this controversy. In ecological speciation with gene flow, the action of diversifying selection (e.g., to adapt to distinct conditions) is opposed by ongoing gene flow that homogenizes gene pools of the nascent species.[3,73–75] The outcome of this selection and gene flow balance depends on (1) the extent of gene flow, (2) the strength of selection, and (3) the linkage disequilibrium (LD) between the target(s) of selection and the "speciation gene(s)" causing incompatibility between the nascent species. Plankton populations in the open ocean critically differ from populations of terrestrial and benthic species with respect to their population size. Although the large population size (N_e) of pelagic species makes selection more powerful,[76] it also results in a higher population-scaled recombination rate, which reduces LD genome-wide. Indeed, the LD was reported to be very low in the *G. huxleyi* genome.[24] This makes it difficult to establish non-random association (that is, LD) between the target(s) of diversifying selection and the genes causing species incompatibility, which is critical for speciation with gene flow to work (e.g., Gavrilets[77] and Flaxman et al.[78]). Thus, the model of speciation with gene flow may not work well in open-ocean plankton, making physical barriers to gene flow essential for speciation. This conclusion is likely applicable to speciation of pelagic organisms beyond phytoplankton. Indeed, in the recent review of speciation with gene flow in marine organisms,[66] only one out of 33 cases was for a pelagic species (cod), which is consistent with the idea that speciation with gene flow is uncommon in pelagic organisms and physical barriers to gene flow play a predominant role in their speciation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Genomic sequencing data
  - Palaeontology data
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Analysis of K-mer spectra
  - Descriptive population genetic analyses
  - Phylogenetic reconstruction
  - Phylogenetic discordances
  - Testing for interspecific hybridization
  - Demographic modeling of speciation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2021.09.073.

ACKNOWLEDGMENTS

We would like to thank Roger Butlin (Univ. Sheffield) for helpful comments on the manuscript, Jeremy Young (UCL) for his guidance with compiling the paleontological dataset, and Ian Probert (Roscoff Culture Collection) for providing strains used in this study. This project has originated from the work funded by Oxford John Fell fund (grant 152/079) to D.A.F. and R.E.M.R. The project received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (APPELS project, grant agreement no. 681746) and the UK Natural Environment Research Council (NERC) award NE/V011049/1. O.A.A. was supported by NERC under grant number NE/S007474/1. R.E.M.R. acknowledges financial support from a Wolfson Research Merit Award.

AUTHOR CONTRIBUTIONS

D.A.F. and R.E.M.R. conceived the study and oversaw the project. D.A.F. participated in culture maintenance and DNA extraction, analyzed the data, and wrote the manuscript. K.H. isolated and characterized the *R. sessilis* strain. O.A.A. compiled the paleontology data. E.M.B. participated in culture maintenance and DNA extraction and contributed phylogenetic analyses. All authors contributed to editing the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 27, 2021
Revised: August 3, 2021
Accepted: September 24, 2021
Published: October 22, 2021

REFERENCES

1. Bowen, B.W., Gaither, M.R., DiBattista, J.D., Iacchei, M., Andrews, K.R., Grant, W.S., Toonen, R.J., and Briggs, J.C. (2016). Comparative phylogeography of the ocean planet. Proc. Natl. Acad. Sci. USA *113*, 7962–7969.
2. Abbott, R., Albach, D., Ansell, S., Arntzen, J.W., Baird, S.J., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C.A., Buggs, R., et al. (2013). Hybridization and speciation. J. Evol. Biol. *26*, 229–246.
3. Coyne, J.A., and Orr, H.A. (2004). Speciation (Sinauer Associates).
4. Miglietta, M.P., Faucci, A., and Santini, F. (2011). Speciation in the sea: overview of the symposium and discussion of future directions. Integr. Comp. Biol. *51*, 449–455.
5. Postel, U., Glemser, B., Salazar Alekseyeva, K., Eggers, S.L., Groth, M., Glöckner, G., John, U., Mock, T., Klemm, K., Valentin, K., and Beszteri, B. (2020). Adaptive divergence across Southern Ocean gradients in the pelagic diatom *Fragilariopsis kerguelensis*. Mol. Ecol. *29*, 4913–4924.
6. Rengefors, K., Kremp, A., Reusch, T.B.H., and Wood, A.M. (2017). Genetic diversity and evolution in eukaryotic phytoplankton: revelations from population genetic studies. J. Plankton Res. *39*, 165–179.
7. Pogson, G.H. (2016). Studying the genetic basis of speciation in high gene flow marine invertebrates. Curr. Zool. *62*, 643–653.
8. Klaveness, D. (1972). *Coccolithus huxleyi* (Lohm.) Kamptn II. The flagellate cell, aberrant cell types, vegetative propagation and life cycles. Br. Phycol. J. *7*, 309–318.
9. Green, J.C., Course, P.A., and Tarran, G.A. (1996). The life-cycle of *Emiliania huxleyi*: a brief review and a study of relative ploidy levels analysed by flow cytometry. J. Mar. Syst. *9*, 33–44.
10. Bendif, E.M., and Young, J. (2014). On the ultrastructure of *Gephyrocapsa oceanica* (Haptophyta) life stages. Cryptogamie, Algologie. *35*, 379–388.
11. Bown, P.R. (1998). Calcareous Nannofossil Biostratigraphy (Kluwer Academic).

12. Raffi, I., Backman, J., Fornaciari, E., Pälike, H., Rio, D., Lourens, L., and Hilgen, F. (2006). A review of calcareous nannofossil astrobiochronology encompassing the past 25 million years. Quat. Sci. Rev. *25*, 3113–3137.
13. Rost, B., and Riebesell, U. (2004). Coccolithophores and the biological pump, responses to environmental changes. In Coccolithophores, from Molecular Process to Global Impact, , H. Thierstein, and J. Young, eds. (Springer-Verlag), pp. 76–99.
14. Henderiks, J., Bartol, M., Pige, N., Karatsolis, B.-T., and Lougheed, B.C. (2020). Shifts in phytoplankton composition and stepwise climate change during the middle Miocene. Paleoceanogr. Paleoclimatol. *35*, e2020PA003915.
15. Milliman, J.D. (1993). Production and accumulation of calcium carbonate in the oceans: budget of a nonsteady state. Global Biogeochem. Cycles *7*, 927–957.
16. Bendif, E.M., Nevado, B., Wong, E.L.Y., Hagino, K., Probert, I., Young, J.R., Rickaby, R.E.M., and Filatov, D.A. (2019). Repeated species radiations in the recent evolution of the key marine phytoplankton lineage *Gephyrocapsa*. Nat. Commun. *10*, 4234.
17. Bendif, M., Probert, I., Díaz-Rosas, F., Thomas, D., van den Engh, G., Young, J.R., and von Dassow, P. (2016). Recent reticulate evolution in the ecologically dominant lineage of Coccolithophores. Front. Microbiol. *7*, 784.
18. Okada, H., and McIntyre, A. (1977). Modern coccolithophores of the Pacific and North Atlantic Oceans. Micropaleontology *23*, 1–55.
19. Frada, M., Young, J., Cachão, M., Lino, S., Martins, A., Narciso, Á., Probert, I., and de Vargas, C. (2010). A guide to extant coccolithophores (Calcihaptophycidae, Haptophyta) using light microscopy. J. Nannoplankton Res. *31*, 58–112.
20. Westbroek, P., Brown, C.W., van Bleijswijk, J., Brownlee, C., Brummer, G.J., Conte, M., Egge, J., Fernández, E., Jordan, R., Knappertsbusch, M., et al. (1993). A model system approach to biological climate forcing. The example of *Emiliania huxleyi*. Global and Planetary Change *8*, 27–46.
21. Paasche, E. (2001). A review of the coccolithophorid *Emiliania huxleyi* (Prymnesiophyceae), with particular reference to growth, coccolith formation, and calcification-photosynthesis interactions. Phycologia *40*, 503–529.
22. Lohbeck, K.T., Riebesell, U., and Reusch, T.B.H. (2012). Adaptive evolution of a key phytoplankton species to ocean acidification. Nat. Geosci. *5*, 346–351.
23. Read, B.A., Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F., Mayer, C., Miller, J., Monier, A., Salamov, A., et al.; Emiliania huxleyi Annotation Consortium (2013). Pan genome of the phytoplankton *Emiliania* underpins its global distribution. Nature *499*, 209–213.
24. Filatov, D.A. (2019). Extreme Lewontin's paradox in ubiquitous marine phytoplankton species. Mol. Biol. Evol. *36*, 4–14.
25. Iglesias-Rodríguez, M.D., Brown, C.W., Doney, S.C., Kleypas, J., Kolber, D., Kolber, Z., Hayes, P.K., and Falkowski, P.G. (2002). Representing key phytoplankton functional groups in ocean carbon cycle models: coccolithophorids. Global Biogeochemical Cycles. *16*, 47-1–47-20.
26. Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics *30*, i541–i548.
27. Simpson, J.T. (2014). Exploring genome characteristics and sequence quality without a reference. Bioinformatics *30*, 1228–1235.
28. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., and Clavijo, B.J. (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics *33*, 574–576.
29. Young, J.R., and Westbroek, P. (1991). Genotypic variation in the coccolithophorid species *Emiliania huxleyi*. Mar. Micropaleontol. *18*, 5–23.
30. Excoffier, L., Smouse, P.E., and Quattro, J.M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics *131*, 479–491.
31. Felsenstein, J. (2006). Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? Mol. Biol. Evol. *23*, 691–700.
32. Pluzhnikov, A., and Donnelly, P. (1996). Optimal sequencing strategies for surveying molecular genetic diversity. Genetics *144*, 1247–1262.
33. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.
34. Kelly, J.K. (1997). A test of neutrality based on interlocus associations. Genetics *146*, 1197–1206.
35. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. *5*, e1000695.
36. Brewer, M.J., Butler, A., and Cooksley, S.L. (2016). The relative performance of AIC, AIC_C and BIC in the presence of unobserved heterogeneity. Methods Ecol. Evol. *7*, 679–692.
37. Krasovec, M., Rickaby, R.E.M., and Filatov, D.A. (2020). Evolution of mutation rate in astronomically large phytoplankton populations. Genome Biol. Evol. *12*, 1051–1059.
38. Wagenmakers, E.J., and Farrell, S. (2004). AIC model selection using Akaike weights. Psychon. Bull. Rev. *11*, 192–196.
39. Martínez-Botí, M.A., Foster, G.L., Chalk, T.B., Rohling, E.J., Sexton, P.F., Lunt, D.J., Pancost, R.D., Badger, M.P.S., and Schmidt, D.N. (2015). Plio-Pleistocene climate sensitivity evaluated using high-resolution CO2 records. Nature *518*, 49–54.
40. Baumann, K.-H., and Freitag, T. (2004). Pleistocene fluctuations in the northern Benguela Current system as revealed by coccolith assemblages. Mar. Micropaleontol. *52*, 195–215.
41. Tangunan, D., Baumann, K.-H., Pätzold, J., Henrich, R., Kucera, M., De Pol-Holz, R., and Groeneveld, J. (2017). Insolation forcing of coccolithophore productivity in the western tropical Indian Ocean over the last two glacial-interglacial cycles. Paleoceanogr. Paleoclimatol. *32*, 692–709.
42. Flores, J.-A., Gersonde, R., Sierro, F.J., and Niebler, H.-S. (2000). Southern Ocean Pleistocene calcareous nannofossil events: calibration with isotope and geomagnetic stratigraphies. Mar. Micropaleontol. *40*, 377–402.
43. Pujos, A., and Giraudeau, J. (1993). Distribution of Noëlaerhabdaceae (calcareous nannofossils) in the Upper and Middle Quaternary of the Atlantic and Pacific oceans. Oceanol. Acta *16*, 349–362.
44. Thierstein, H.R., Geitzenauer, K.R., Molfino, B., and Shackleton, N.J. (1977). Global synchroneity of late Quaternary coccolith datum levels - validation by oxygen isotopes. Geology *5*, 400–404.
45. Gard, G., and Crux, J.A. (1991). Preliminary results from Hole 704A: Arctic–Antarctic correlation through nannofossil biochronology. Proc. Ocean Drilling Progr. Sci. Results *114*, 193–200.
46. Gard, G. (1989). Variations in coccolith assemblages during the last glacial cycle in the high and mid-latitude Atlantic and Indian Oceans. In Nannofossils and Their Applications, , J.A. Crux, and S.E. Heck, eds. (Ellis Horwood), pp. 108–121.
47. Novacyzk, N.R., and Baumann, M. (1992). Combined high-resolution magnetostratigraphy and nannofossil biostratigraphy for late Quaternary Arctic Ocean sediments. Deep-Sea Res. *39*, S567–S601.
48. Gartner, S. (1977). Calcareous nannofossil biostratigraphy and revised zonation of the Pleistocene. Mar. Micropaleontol. *2*, 1–25.
49. Emiliani, C. (1993). Extinction and viruses. Biosystems *31*, 155–159.
50. Crow, J.F., and Kimura, M. (1970). An Introduction to Population Genetics Theory (Harper & Row).
51. Vucetich, J.A., Waite, T.A., and Nunney, L. (1997). Fluctuating population size and the ratio of effective to census population size. Evolution *51*, 2017–2021.
52. Durand, E.Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. Mol. Biol. Evol. *28*, 2239–2252.

53. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–1093.

54. Matsuoka, H., and Okada, H. (1989). Quantitative analysis of Quaternary nannoplankton in the subtropical northwestern Pacific Ocean. Mar. Micropaleontol. *14*, 97–118.

55. Wei, W. (1993). Calibration of upper Pliocene-lower Pleistocene nannofossil events with oxygen isotope stratigraphy. Paleoceanogr. Paleoclimatol. *8*, 85–99.

56. Hewitt, G.M. (2004). Genetic consequences of climatic oscillations in the Quaternary. Philos. Trans. R. Soc. Lond. B Biol. Sci. *359*, 183–195, discussion 195.

57. Laakkonen, H.M., Hardman, M., Strelkov, P., and Väinölä, R. (2021). Cycles of trans-Arctic dispersal and vicariance, and diversification of the amphi-boreal marine fauna. J. Evol. Biol. *34*, 73–96.

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

59. Sato, T., and Kameo, K. (1996). Pliocene to Quaternary calcareous nannofossil biostratigraphy of the Arctic Ocean with reference to late Pliocene glaciation. Proc. Ocean Drilling Prog. Scientific Results *151*, 39–58.

60. Takayama, T., and Sato, T. (1987). Coccolith biostratigraphy of the North Atlantic Ocean, deep sea drilling project leg 94. Initial Rep. Deep Sea Drill. Proj. *94*, 651–702.

61. Lazarus, D. (1994). Neptune: a marine micropaleontology database. Math. Geol. *26*, 817–832.

62. Jablonski, D. (1993). The tropics as a source of evolutionary novelty through geological time. Nature *364*, 142–144.

63. Rolland, J., Condamine, F.L., Jiguet, F., and Morlon, H. (2014). Faster speciation and reduced extinction in the tropics contribute to the Mammalian latitudinal diversity gradient. PLoS Biol. *12*, e1001775.

64. Hellberg, M.E. (2009). Gene flow and isolation among populations of marine animals. Annu. Rev. Ecol. Evol. Syst. *40*, 291–310.

65. Faria, R., Johannesson, K., and Stankowski, S. (2021). Speciation in marine environments: Diving under the surface. J. Evol. Biol. *34*, 4–15.

66. Potkamp, G., and Fransen, C.H.J.M. (2019). Speciation with gene flow in marine systems. Contrib. Zool. *88*, 133–172.

67. Hawkings, J.R., Wadham, J.L., Tranter, M., Raiswell, R., Benning, L.G., Statham, P.J., Tedstone, A., Nienow, P., Lee, K., and Telling, J. (2014). Ice sheets as a significant source of highly reactive nanoparticulate iron to the oceans. Nat. Commun. *5*, 3929.

68. Hawkings, J., Wadham, J., Tranter, M., Telling, J., Bagshaw, E., Beaton, A., Simmons, S.-L., Chandler, D., Tedstone, A., and Nienow, P. (2016). The Greenland Ice Sheet as a hot spot of phosphorus weathering and export in the Arctic. Global Biogeochem. Cycles *30*, 191–210.

69. Arrigo, K.R., van Dijken, G.L., Castelao, R.M., Luo, H., Rennermalm, Å.K., Tedesco, M., Mote, T.L., Oliver, H., and Yager, P.L. (2017). Melting glaciers stimulate large summer phytoplankton blooms in southwest Greenland waters. Geophys. Res. Lett. *44*, 6278–6285.

70. Kasten, S., Haese, R.R., Zabel, M., Rühlemann, C., and Schulz, H.D. (2001). Barium peaks at glacial terminations in sediments of the equatorial Atlantic Ocean - relicts of deglacial productivity pulses? Chem. Geol. *175*, 635–651.

71. Palumbi, S.R. (1994). Genetic divergence, reproductive isolation, and marine speciation. Annu. Rev. Ecol. Syst. *25*, 547–572.

72. Bowen, B.W., Rocha, L.A., Toonen, R.J., and Karl, S.A.; ToBo Laboratory (2013). The origins of tropical marine biodiversity. Trends Ecol. Evol. *28*, 359–366.

73. Rundle, H.D., and Nosil, P. (2005). Ecological speciation. Ecol. Lett. *8*, 336–352.

74. Nosil, P. (2012). Ecological Speciation (Oxford University).

75. Hendry, A.P., Nosil, P., and Rieseberg, L.H. (2007). The speed of ecological speciation. Funct. Ecol. *21*, 455–464.

76. Peijnenburg, K.T.C.A., and Goetze, E. (2013). High evolutionary potential of marine zooplankton. Ecol. Evol. *3*, 2765–2781.

77. Gavrilets, S. (2006). The Maynard Smith model of sympatric speciation. J. Theor. Biol. *239*, 172–182.

78. Flaxman, S.M., Feder, J.L., and Nosil, P. (2012). Spatially explicit models of divergence and genome hitchhiking. J. Evol. Biol. *25*, 2633–2650.

79. von Dassow, P., John, U., Ogata, H., Probert, I., Bendif, M., Kegel, J.U., Audic, S., Wincker, P., Da Silva, C., Claverie, J.-M., et al. (2015). Life-cycle modification in open oceans accounts for genome variability in a cosmopolitan phytoplankton. ISME J. *9*, 1365–1377.

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

81. Song, L., Florea, L., and Langmead, B. (2014). Lighter: fast and memory-efficient sequencing error correction without counting. Genome Biol. *15*, 509.

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

83. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

84. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

85. Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. *78*, 629–644.

86. Ramos-Onsins, S.E., Ferretti, L., Raineri, E., Jené, J., Marmorini, G., Burgos, W., and Vera, G. (2018). mstatspop: statistical analysis using multiple populations for genomic data. https://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/software.html.

87. Malinsky, M., Matschiner, M., and Svardal, H. (2021). Dsuite - Fast D-statistics and related admixture evidence from VCF files. Mol. Ecol. Resour. *21*, 584–595.

88. Filatov, D.A. (2009). Processing and population genetic analysis of multigenic datasets with ProSeq3 software. Bioinformatics *25*, 3189–3190.

89. Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. *35*, 1547–1549.

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

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

92. Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics *20*, 289–290.

93. Wickham, H. (2009). Ggplot2: Elegant Graphics for Data Analysis (Springer).

94. Wilke, C.O. (2021). cowplot – Streamlined plot theme and plot annotations for ggplot2. https://wilkelab.org/cowplot/index.html.

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

96. Schlitzer, R. (2021). Ocean data view. https://odv.awi.de.

97. Keller, M.D., Selvin, R.C., Claus, W., and Guillard, R.R.L. (1987). Media for the culture of oceanic ultraphytoplankton. J. Phycol. *23*, 633–638.

98. Nei, M. (1987). Molecular Evolutionary Genetics (Columbia University).

99. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

100. Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. *10*, 512–526.

Please cite this article in press as: Filatov et al., The mode of speciation during a recent radiation in open-ocean phytoplankton, Current Biology (2021), https://doi.org/10.1016/j.cub.2021.09.073

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
Illumina high throughput sequencing	WTCHG, Oxford	N/A
QIAGEN DNEASY Plant Kit	QIAGEN	cat # 69104
Deposited data		
WGS data	This paper	NCBI Bioproject: PRJNA532411
Stratigraphic dataset	This paper	https://doi.pangaea.de/10.1594/PANGAEA.935786
Experimental models: Organisms/strains		
Gephyrocapsa huxleyi A	Krasovec et al.[37]	CCMP1516; SAMN13932591
G. huxleyi A	Bendif et al.[16]	ARC30-1; SRR8885247
G. huxleyi A	von Dassow et al.[79]	RCC4002; ERR695588
G. huxleyi A	von Dassow et al.[79]	RCC4028; ERR695589
G. huxleyi A	von Dassow et al.[79]	RCC4030; ERR695590
G. huxleyi A	this paper	PLY829; SRR14251552
G. huxleyi A	this paper	PLY853; SRR14251551
G. huxleyi A	this paper	OA15; SRR14251540
G. huxleyi A	this paper	SO21-1; SRR14251531
G. huxleyi A	this paper	ARC27-1; SRR14251530
G. huxleyi A	this paper	ARC39-1; SRR14251529
Gephyrocapsa huxleyi B	this paper	ARC68-2; SRR14251528
G. huxleyi B	this paper	RCC5137; SRR14251527
G. huxleyi B	this paper	RCC5141; SRR14251526
G. huxleyi B	this paper	RCC1239; SRR14251525
G. huxleyi B	this paper	RCC5134; SRR14251550
G. huxleyi B	this paper	RCC6566; SRR14251549
G. huxleyi B	this paper	RCC1212; SRR14251548
G. huxleyi B	Bendif et al.[16]	RCC1253; SRR8885248
G. huxleyi B	Read et al.[23]	van556; SRR391483
G. huxleyi B	Read et al.[23]	PLY92E; SRR391484
G. huxleyi B	Read et al.[23]	PLY92D; SRR391472
Gephyrocapsa muellerae	Bendif et al.[16]	RCC3370; SRR8885246
G. muellerae	this paper	RCC3898; SRR14251547
G. muellerae	this paper	RCC3862; SRR14251546
G. muellerae	this paper	RCC5119; SRR14251545
Gephyrocapsa parvula	Bendif et al.[16]	RCC4033; SRR8885241
G. parvula	Bendif et al.[16]	RCC4034; SRR8885242
G. parvula	this paper	RCC4035; SRR14251544
G. parvula	this paper	RCC4036; SRR14251543
Gephyrocapsa ericsonii	Bendif et al.[16]	RCC4032; SRR8885244
Gephyrocapsa oceanica	this paper	BOUM49; SRR14251541
G. oceanica	this paper	OA18; SRR14251539
G. oceanica	this paper	RCC1238; SRR14251538
G. oceanica	this paper	RCC1281; SRR14251537
G. oceanica	Bendif et al.[16]	RCC1296; SRR8885245
G. oceanica	this paper	RCC1310; SRR14251536
G. oceanica	this paper	RCC1314; SRR14251535

(*Continued on next page*)

Please cite this article in press as: Filatov et al., The mode of speciation during a recent radiation in open-ocean phytoplankton, Current Biology (2021), https://doi.org/10.1016/j.cub.2021.09.073

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
G. oceanica	this paper	RCC1562; SRR14251534
G. oceanica	this paper	RCC1836; SRR14251533
G. oceanica	Bendif et al.[16]	RCC3711; SRR8885243
G. oceanica	this paper	RCC3733; SRR14251532
Reticulofenestra sessilis	this study	RCC6730; SRR14251542
Software and algorithms		
Trimmomatic v.0.30	Bolger et al.[80]	http://www.usadellab.org/cms/?page=trimmomatic
Kmer Analysis Toolkit v2.4.1	Mapleson et al.[28]	https://kat.readthedocs.io/en/latest/
Bowtie2 v. 2.1.0	Song et al.[81]	http://bowtie-bio.sourceforge.net/bowtie2/
bwa v0.7	Li and Durbin[82]	http://bio-bwa.sourceforge.net/bwa.shtml
GATK v3.4	McKenna et al.[83]	https://gatk.broadinstitute.org/hc/en-us
Samtools v1.7	Li et al.[84]	http://samtools.sourceforge.net/
fastPHASE v1.4.0	Scheet and Stephens[85]	http://scheet.org/software.html
mstatspop v.0.1b	Ramos-Onsins et al.[86]	https://github.com/CRAGENOMICA/mstatspop
δaδI v2.1.0	Gutenkunst et al.[35]	https://bitbucket.org/gutenkunstlab/dadi/src/master
Python code defining IM, SC, AM and VM models	This paper	https://sourceforge.net/projects/rundadi/
Custom ABBA/BABA script	Bendif et al.[16]	https://github.com/brunonevado/calcD_from_fas
Dsuite v0.4 r38	Malinsky et al.[87]	https://github.com/millanek/Dsuite
ProSeq3 v3.994	Filatov[88]	https://sourceforge.net/projects/proseq/
MEGA-X v10.1.5	Kumar et al.[89]	https://www.megasoftware.net/
RAxML v8.2.12	Stamatakis[90]	https://cme.h-its.org/exelixis/web/software/raxml/
R package phangorn v2.7.1	Schliep[91]	https://cran.r-project.org/web/packages/phangorn/index.html
R package ape v5.5	Paradis et al.[92]	https://cran.r-project.org/web/packages/ape/index.html
R package ggplot2	Wickham[93]	https://ggplot2.tidyverse.org/
R package cowplot	Wilke[94]	https://wilkelab.org/cowplot/index.html
Densitree v2.2.5	Bouckaert[95]	https://www.cs.auckland.ac.nz/~remco/DensiTree/
ASTRAL v5.7.1	Mirarab et al.[26]	https://github.com/smirarab/ASTRAL
Ocean Data View v5.5	Schlitzer[96]	https://odv.awi.de

RESOURCE AVAILABILITY

Lead contact

Further information regarding the manuscript and requests for reagents may be directed to, and will be fulfilled by the lead contact, Dmitry A. Filatov (Dmitry.Filatov@plants.ox.ac.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

High throughput sequencing data generated in this paper is available from NCBI under bioproject number PRJNA532411. Accession numbers are listed in the Table S1 and the key resources table. The cleaned stratigraphic dataset used in the paper is available from the repository Pangaea (https://doi.pangaea.de/10.1594/PANGAEA.935786).

All original code has been deposited at Sourceforge.net and is publicly available as of the date of publication using the following URL: https://sourceforge.net/projects/rundadi/

Any additional information required to re-analyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Twenty-eight Noëlaerhabdaceae strains that were newly sequenced in this study (listed in Table S1). The cultures from twenty-seven *Gephyrocapsa* strains were obtained from Roscoff culture collection. These include six strains of *G. huxleyi* A, seven strains of *G. huxleyi* B, nine *G. oceanica* strains, three *G. muellerae* strains and two strains of *G. parvula*. The *R. sessilis* strain we analyzed

Please cite this article in press as: Filatov et al., The mode of speciation during a recent radiation in open-ocean phytoplankton, Current Biology (2021), https://doi.org/10.1016/j.cub.2021.09.073

was isolated and provided by co-author Kyoko Hagino. Prior to genomic DNA extraction these clonal strains were maintained in K/2(-Si,-Tris,-Cu) medium[97] at 17°C with 50 μmol-photons.m^{-2}.s^{-1} illumination provided by daylight neon tubes with a 14:10h L:D cycle.

METHOD DETAILS

Genomic sequencing data

The sources of genome sequence data used in the paper are summarized in Table S1. For 28 strains newly sequenced in this study the cell cultures were harvested by centrifugation (4500 g, 10 min), washed twice with TE buffer, and DNA was extracted with Plant DNAeasy QIAGEN kit. For each sample, quantifications of nucleic acids were performed either with a Qubit 3.0 fluorometer (Thermofisher Scientific) and a Nanodrop. Total DNA extracts were sequenced on Illumina platform at the Wellcome Trust Centre for Human Genomics, Oxford. Paired-end libraries were prepared individually, barcoded, and then combined prior to sequencing. Libraries were sequenced using an Illumina NovaSeq instrument to produce 150 base-pair (bp) paired-end reads. After quality trimming with Trimmomatic,[80] the sequence reads were mapped to *G. huxleyi* reference genome[23] with *bwa* v0.7.[82] Relatively low sequence divergence between the strains (see Table 1) resulted in high (> 50%) proportion of reads mapping to the reference. Duplicate reads were removed with *samtools* v1.2 (https://samtools.sourceforge.net), and regions around indels were realigned with *GATK* v3.4.[83] SNP calling was done for all strains simultaneously, with *samtools* and *bcftools* v 1.2 (part of *samtools* package) using the alternative multiallelic variant caller (-m option) and including homozygous blocks with minimum depth of 8 (-g 8), after excluding reads with mapping quality below 20 (-q 20) and bases with base quality below 20 (-Q 20). SNPs with fewer than 8 reads supporting it, within 3 bp of an indel, with quality below 10, or with fewer than 2 reads supporting each allele (for heterozygous calls) were marked as low-quality and excluded from further analysis. Resulting multisample *vcf.* files including confident SNP calls and homozygous blocks were imported to proSeq3 software,[88] which was used to convert the data to fasta, mega and dadi formats.

Palaeontology data

In order to assess patterns of diachroneity in the marine realm and constrain its possible causes, we have reviewed the literature and compiled the first occurrences of *G. huxleyi* and other species of *Gephyrocapsa*, as revealed from the sediment record (Table S6). This new synthesis includes previously published data, all with rigorous relationships of their emergence events with marine isotope stages based on good quality oxygen isotope stratigraphy and/or astronomical tuning from each sediment core. To ensure reliability of the pattern observed and to account for possible inconsistency of age models in individual cores we included only the areas with at least two cores analyzed in the same area. Data processing and visualization (Figure 3A) was completed using R packages ggplot2[93] and cowplot.[94] The spatial distribution of sediment cores (Figure 3B) was generated using Ocean Data View software.[96]

QUANTIFICATION AND STATISTICAL ANALYSIS

Analysis of K-mer spectra

The analysis of K-mer spectra was conducted with hist function in the Kmer Analysis Toolkit (KAT) V2.4.1.[28] K-mer spectra show how many fixed length words (k-mers) appear a certain number of times in the sequence data. The frequency of occurrence is plotted on the x axis and the number of k-mers on the y axis. K-mer spectrum allows one to gain insight into the genome ploidy without genome assembly, with the analysis done on raw sequence reads. K-mer is a 'word' of k nucleotides long; in this analysis we used k = 23, but k = 19 and k = 27 give very similar results (data not shown). One can generate a list of k-mers that can be found in sequence data, record for each k-mer how many times it is seen in the data and build a histogram. For a haploid genome the histogram is expected to form a unimodal distribution, while for higher ploidies more peaks should be observed.[27,28] For a diploid genome two peaks should be observed, with the right peak comprising k-mers corresponding to sequence content that is identical in two copies of the diploid genome, while the left peak includes k-mers corresponding to unique sequence content, such as heterozygous sites that distinguish the copies of the diploid genome. K-mer spectra for all species analyzed in this study formed two peaks, indicating diploidy, as shown for *G. muellerae*, *G. oceanica* and *G. huxleyi* sequences on Figure S1.

Descriptive population genetic analyses

The annotations for the coding regions (CDS) from file "Emihu1_best_genes.gff" available for the reference genome,[23] were used to identify silent and non-silent sites in polymorphism analyses. Average heterozygosity (π) at different types of sites was calculated using mstatspop.[86] The same software was also used to calculate a range of summary statistics for intraspecific polymorphism (Tajima's D,[33] Z_{nS}[34]) and interspecific differentiation and divergence (F_{st}, D_{xy}[98]).

Phylogenetic reconstruction

For phylogenetic analysis of the Noëlaerhabdaceae, best-covered and aligned contigs were filtered out retaining 243 contigs with no missing individual and around 80% positions completeness. For each contig, haplotypes were inferred using fastPHASE v1.4.0 and one haplotype was selected per individual. Gene tree topologies were then assessed through a stepping window approach, contig-alignments were split in 910 separate regions 10 kb long separated by at least 25 kb to minimize linkage between subsequent windows, after excluding the alignment positions with gaps or missing data. For each of these 910 alignments, we performed a

Please cite this article in press as: Filatov et al., The mode of speciation during a recent radiation in open-ocean phytoplankton, Current Biology (2021), https://doi.org/10.1016/j.cub.2021.09.073

phylogenetic reconstruction using the GTRGAMMA model and 100 bootstrap replicates in RAxML. Best ML trees with bootstrap replicates were then used to produce a multicoalescent species tree using ASTRAL.[26]

The chloroplast phylogeny was reconstructed from consensus contigs called with samtools mpileup command for two strains per species (Figure S3) except *R. sessilis* and *G. ericsonii* for which only one strain per species is available. To create the consensus for each of these *Gephyrocapsa* and *R. sessilis* strains, we used short reads that were mapped to the *G. huxleyi* reference chloroplast genome (NCBI accession JN022705.1) as part of sequence read mapping described in "Genomic sequencing data" section above. The resulting consensus sequences were aligned to chloroplast genome sequences of outgroups *Tisochrysis lutea* (NCBI accession NC_040291.1) and *Isochrysis galbana* (NCBI accession MT304829.1) using muscle software.[99] The alignments were converted to mega format with ProSeq3 software[88] and used for reconstruction of the phylogeny with the maximum likelihood method and Tamura-Nei model[100] as implemented in MEGA software.[89]

Phylogenetic discordances

A DensiTree plot (left side of Figure 1) was produced using Densitree[95] version 2.2.1 to visualize phylogenetic discordances between loci based on 910 ML trees reconstructed for 10 kb long genomic regions. Each of the 910 genomic regions was used separately (without concatenation) to reconstruct region-specific phylogenies using RAxML, as described in the previous section. The total number of sites used in this analysis was 9,100,000. For the phylogenies reconstructed for each of these 10 kb fragments, we used the pruneTree function in the R phangorn package[91] to collapse nodes with bootstrap support < 75%. Trees with no nodes over 75% bootstrap support were discarded. Using the root function, each of the pruned trees was then rooted by *R. sessilis*, and each tree was made ultrametric using the chronos function with default settings in the R ape package.[92] Resulting trees were then loaded into DensiTree software[95] that was used to generate the densiTree plot (Figure 1A). The consensus tree was produced with DensiTree software with the following settings: star tree, consensus width = 1, consensus intensity 28.1, and default values for all other settings.

Testing for interspecific hybridization

We performed Patterson's[52,53] *D*-statistic and f_4-ratio tests (Table S5; Figure S4) which compare two phylogenetically incongruent site patterns of ancestral (A) and derived (B) alleles ABBA—(((A,B),B),A) and BABA—(((B,A),B),A) on a four-taxon phylogeny with the topology: (((P1,P2),P3),Outgroup). If the incongruence is due to incomplete lineage sorting, 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. Z-scores and p values were then used to determine significance. We used a custom ABBA/BABA script (available at https://github.com/brunonevado/calcD_from_fas) and Dsuite[87] to test every phylogenetically congruent three-species subtree using *R. sessilis* as outgroup. In the Dsuite package, *D*-statistics are always positive as P1 and P2 are ordered along with the f_4-ratio statisitic. Z-scores and associated *P*-values were calculated by block-jackknife procedure[52] to assess the significance for a deviation of the *D*-statistic from zero. Bonferroni-adjusted *P*-value < 0.05 indicated potential signal of gene flow.

Demographic modeling of speciation

For demographic inference (Figures 2A and 2D; Tables 3, S3, and S4) we used the δaδi package[35] with pairs and trios of species. The analyses involving trios of species did not converge even for simple models, hence we restricted the analyses for pairs of species, using multiple species combinations per species split, whenever possible. For all pairwise analyses we used 'unfolded' 2-dimensional site frequency spectrum, where ancestral and derived alleles were identified using outgroups (*G. oceanica* + *R. sessilis* for all other species and *R. sessilis* for the analyses involving *G. oceanica*). After filtering out the SNPs with missing data, the dataset used in the analysis included 321,327 SNPs.

The models used in the analysis are summarized in Table S3 and the python code defining and running these models is available: https://sourceforge.net/projects/rundadi/. To ensure efficient use of multicore processors, the code implements parallel running of multiple instances of model optimization with different starting parameters, which maximizes the chance of finding the global maximum. In our analyses, each model was re-run at least 40 times with different starting parameters (via ".perturb_params" function in δaδi). The code also implements parametric bootstrap (via ".sample" function in δaδi), which was used to estimate the bootstrap confidence intervals for the parameter estimates for the best-fitting models, based on 100 bootstrap replicates (Table 3).

As our initial analyses revealed that allowing for interspecific gene flow and population size change significantly improve the fit of models to data, the speciation models we used in the paper include both population size change and interspecific migration rate that is either constant (isolation-migration [IM] models) or variable (IM-eM, ancestral migration [AM] and secondary contact [SC] models) over time (Table S3). Furthermore, the IM, AM and SC models were modified by adding additional parameters to improve fit to data. In particular, the "m2" models (IMm2, AMm2 and SCm2) allowed for different migration rates in two directions, while the "hm" models (e.g., IMm1_hm) accounted for heterogeneity of migration rate across the genome with *p* and 1-*p* proportions of the genome having different migration rates. Furthermore, possible heterogeneity of effective population size across the genome (e.g., between genomic regions with frequent and rare recombination) was taken into account in the "hn" models (e.g., IMm1_hn). The combinations of "hm" and "hn" modes ("hmhn" and "hm2hn") were also analyzed (e.g., IMm1_hm2hn; see Tables S3 and S4). The fit of models to data was assessed with log-likelihood. Model fit to data was compared using the sample size corrected Akaike information criterion (AICc[36]) to

Please cite this article in press as: Filatov et al., The mode of speciation during a recent radiation in open-ocean phytoplankton, Current Biology (2021), https://doi.org/10.1016/j.cub.2021.09.073

rank the different models. The best fitting model for the given pair of species was used as a reference to calculate the Akaike weights (formula 4 in Wagenmakers et al.;[38] Table S4) that can be interpreted as the probability that the model is the best.

The parameter estimates are initially in units of ancestral population size (N_a). To calculate the N_a and to convert the parameter estimates to biologically meaningful values we used the value of theta estimated by the program and the mutation rate (μ = $5.55x10^{-10}$) measured for *G. huxleyi* in a mutation accumulation experiment.[37] For confidence intervals of parameter estimates [in brackets in Table 3] we conservatively used the minimal and maximal estimates for 100 bootstrap replicates, taking the uncertainty of mutation rate estimate (m = 5.05 to 6.09 $x10^{-10}$ per nucleotide per cell division[37]) into account. That is, the minimal and maximal parameter estimates were calculated assuming the mutation rates $m = 6.09*10^{-10}$ and $5.05*10^{-10}$, respectively. The population size estimates for each species were multiplied by N_a to calculate the size in units of individuals. The times of speciation (and other times – of secondary contact, T_{SC} etc.) were converted into units of generations by multiplying the estimates by $2N_a$. Finally, to convert the number of generations into years we assumed 50 generations per year. Under optimal growth conditions in lab culture the *G. huxleyi* generation time is 1.17 days, on average,[37] but given that growth conditions in nature vary through the year, the average time between cell divisions in nature is likely much longer than one day. Furthermore, the frequency of meiotic versus mitotic cell divisions is not known, which makes it difficult to be certain about the actual number of generations per year. Assuming 50 generations per year, on average, appears biologically realistic and yields the correct timing for the origin of *G. huxleyi* – the event well documented in the fossil record,[12] that we implicitly use as a calibration point for the "generations per year" parameter.

The number of migration parameters per model ranged from 0 to 4 (Table S3), with some models allowing for separate migration rates in two directions and for variation of migration across the genome. To obtain the mean migration rate listed in Table 3, we averaged all estimates of migration per model per species pair, taking into account the proportion of the genome for which the particular migration rate was estimated. That is, for the "hm" models that allow for different migration at two classes of sites in the genome, the average migration was calculated as $m_{avrg} = (m_a p + m_b(1-p))/2$, where m_a and m_b are estimates of migration for sites classes A and B, while p is the proportion of the genome estimated to belong to site class A, with the rest belonging to site class B.