scientific reports

OPEN

Empirical evidence for concerted evolution in the 18S rDNA region of the planktonic diatom genus *Chaetoceros*

Daniele De Luca[1,3]✉, Wiebe H. C. F. Kooistra[1], Diana Sarno[2], Elio Biffali[2] & Roberta Piredda[1]✉

Concerted evolution is a process of homogenisation of repetitive sequences within a genome through unequal crossing over and gene conversion. This homogenisation is never fully achieved because mutations always create new variants. Classically, concerted evolution has been detected as "noise" in electropherograms and these variants have been characterised through cloning and sequencing of subsamples of amplified products. However, this approach limits the number of detectable variants and provides no information about the abundance of each variant. In this study, we investigated concerted evolution by using environmental time-series metabarcoding data, single strain high-throughput sequencing (HTS) and a collection of Sanger reference barcode sequences. We used six species of the marine planktonic diatom genus *Chaetoceros* as study system. Abundance plots obtained from environmental metabarcoding and single strain HTS showed the presence of a haplotype far more abundant than all the others (the "dominant" haplotype) and identical to the reference sequences of that species obtained with Sanger sequencing. This distribution fitted best with Zipf's law among the rank abundance/ dominance models tested. Furthermore, in each strain 99% of reads showed a similarity of 99% with the dominant haplotype, confirming the efficiency of the homogenisation mechanism of concerted evolution. We also demonstrated that minor haplotypes found in the environmental samples are not only technical artefacts, but mostly intragenomic variation generated by incomplete homogenisation. Finally, we showed that concerted evolution can be visualised inferring phylogenetic networks from environmental data. In conclusion, our study provides an important contribution to the understanding of concerted evolution and to the interpretation of DNA barcoding and metabarcoding data based on multigene family markers.

The first DNA hybridisation studies conducted in the mid-1960–1970s showed that a large fraction of eukaryotic genomes was composed of repetitive regions[1,2]. When comparing the repetitive DNA families, a greater sequence similarity within species than between them was observed[3,4]. Such observation was incompatible with the common model of divergent evolution, according to which the differences in nucleotide sequence between different repeats of the same species were expected to be as large as those between repeats of different species[5]. Therefore, there had to be a mechanism responsible for the homogenisation of such sequences. The expression "concerted evolution"[6] was coined to indicate this phenomenon, by which an individual member of a gene family evolves in the same (concerted) way as all the other members of the family[7].

The best-known example of concerted evolution are the ribosomal DNA cistrons (rDNA)[8,9], but also other non-coding regions and genes (e.g. globins, immunoglobulins, heat-shock genes, histones) are known to evolve in this way[8,10–12]. Two processes, gene conversion and unequal crossing-over, will eventually lead to sequence homogeneity in absence of mutations[13–15]. However, mutations always occur and gene conversion and unequal crossing-over also contribute to their spread, generating variation across repeats[5]. These deviations from sequence homogenisation have been detected in animals[16,17], fungi[18], protists[19] and especially in plants[20–24]. The extent of such deviations is important in the case of the rDNA cistron, since it is the target for DNA barcoding studies in

[1]Department of Integrative Marine Ecology, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [2]Department of Research Infrastructure for Marine Biological Resources, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [3]Present address: Department of Biology, University of Naples Federico II, Botanical Garden of Naples, Via Foria 223, 80139 Naples, Italy. ✉email: daniele.deluca088@gmail.com; robpiredda@gmail.com

such taxa as fungi and protists[25–27]. Therefore, understanding the inheritance of rDNA genes and spacers is vital for taxonomic and systematic studies involving them.

So far, intra-individual variation among the rDNA-cistrons has been spotted as noise in electropherograms and as variation among sequenced clones generated from the PCR products[28,29]. The resulting sequences, together with ones from closely related species, were then analysed phylogenetically to ascertain the degree of relatedness within and among species[22,30–32]. However, this approach has two main limitations: the number of detectable variants is constrained by the number of sequenced clones and there is no information about the abundance of each variant. In recent times, concerted evolution has also been revealed by whole-genome shotgun sequence data[9] and chromosomal and array approaches[33,34]. Nowadays, the use of the high-throughput sequencing approach, which allows sequencing of thousands of copies of a target region from environmental samples and bulk communities (metabarcoding) and even single specimens (single strain HTS), could overcome the limitations associated with cloning and complexity of analysis of whole-genome shotgun data[35,36]. Therefore, metabarcoding can be particularly useful to study concerted evolution. A recent work by[37] assessed the diversity in the marine planktonic diatom family Chaetocerotaceae through a 18S-V4 rDNA environmental metabarcode haplotype collection generated from 48 seasonal samples taken at the long-term ecological research station (LTER) MareChiara (MC) in the Gulf of Naples, Mediterranean Sea, Italy. A phylogeny inferred from the chaetocerotacean metabarcode haplotypes and (Sanger-sequenced) taxonomic references showed the following recurrent patterns. Terminal clades comprised from just a few to several hundreds of haplotypes, one of which was far more abundant than the others. That one, called "dominant haplotype", was identical or nearly identical to the Sanger-sequence of a reference strain, when available, of that clade[37]. Furthermore, the relationship among "dominant" and "minor" haplotypes was consistent across species: when plotted on a logarithmic scale, the reads of the dominant haplotype were at least one order more abundant than those of any of other haplotypes of that species. Based on that signal, the authors hypothesized that such a pattern "result from an equilibrium between the appearance of novel haplotypes, random drift, and the homogenizing effect of concerted evolution"[37].

In the present study, we designed an experiment and generated new data to confirm the hypothesis of concerted evolution in the 18S rDNA gene of several *Chaetoceros* species by high-throughput sequencing (HTS) of the V4 region of monoclonal strains. In particular, we used the new data obtained from monoclonal strains together with the previous environmental chaetocerotacean metabarcode dataset and the *Chaetoceros* reference barcodes obtained with Sanger sequencing to test if: (i) the most abundant haplotype in each single strain matches the reference barcode of that strain obtained with Sanger sequencing and with the dominant environmental haplotype; (ii) the minor haplotypes found in the environmental metabarcoding data are present in the strains of each species; (iii) the temporal distribution of dominant and minor haplotypes in environmental data is the same.

Methods

Selection of taxa to study concerted evolution. In order to test the aforementioned hypotheses, we used the metabarcoding data of Chaetocerotaceae from the LTER MareChiara (Gulf of Naples) from[37] deposited in GenBank at the accession numbers MK938374–MK940235 (414,041 reads). The dataset includes sequences gathered from 48 dates across three years (2011–2013); sampling procedure is described in detail in[38]. In this study, the term "haplotype" indicates the non-redundant (unique) sequences. Based on the chaetocerotacean metabarcode haplotype diversity illustrated in the phylogeny presented in[37], we selected for detailed HTS analysis strains of six species representing different clades of the family tree and showing different read distribution patterns over the environmental haplotypes or over the seasonal cycle. In particular, we chose: *Chaetoceros tenuissimus* as representative of species occurring at high abundance all over the year and displaying many minor haplotypes; *C. costatus* as species with a marked seasonality, displaying also a few minor haplotypes at high abundances; *C. anastomosans* as a relatively rare species, displaying a single, lowly abundant, dominant haplotype in the environmental data; *C. curvisetus* 2 as species common all over the year with few minor haplotypes; *Chaetoceros* spp. Na11C3 and Na26B1 as examples of two closely related species, i.e., with distinct reference barcodes and dominant haplotypes, but recovered together in an internally unresolved clade with minor haplotypes. For each species, we selected outgroup taxa (Supplementary Table S1) for subsequent validation of sequences gathered from BLAST analysis.

Analysis of environmental sequences. As time-series data we used V4 metabarcode reads generated from 48 plankton samples taken at the Long Term Ecological Research (LTER) station MareChiara in the Gulf of Naples (Mediterranean Sea) over the seasonal cycles of three consecutive years (2011–2013)[37,38]. To retrieve sequences of the selected *Chaetoceros* species in the chaetocerotacean dataset, we used the respective 18S reference sequences and close outgroups as queries for a local BLAST[39] at 95%. The metabarcodes extracted were then aligned with the references and the outgroup taxa using MAFFT online[40] and a phylogenetic tree was built in FastTree v2.1.8[41], using the GTR model. The resulting tree was visualised and modified in Archaeopteryx v0.9901[42] in order to remove sequences clustering within outgroup clades and gather only metabarcodes of the species of interest. The sequences retrieved were considered validated and used to retrieve the info of abundance using mothur v1.41.1[43].

Single strain HTS. Single strain metabarcoding was performed on: two strains of *C. anastomosans*, four strains of *C. costatus*, four strains of *C. curvisetus* 2, one of *Chaetoceros* sp. Na26B1, two of *Chaetoceros* sp. Na11C3 and three strains of *C. tenuissimus* (Table 1). Strains were obtained from cell chains collected at the LTER MareChiara. For each sample, we performed individual PCR in two steps: a first reaction for the amplification of the target sequence, and a second reaction (using the PCR product of the former one as template) to ligate proprietary adaptor sequence (P1) and unique 10–12 bp long identifier nucleotide key tags (barcodes) com-

Species	Strain
C. anastomosans	Na14C2
	Na14C3
C. costatus	Na1A3
	Na32B1
	Ro1B1
	Ro2A2
C. curvisetus 2	Ch5B2
	Na1C1
	Na19A2
	Na20A4
Chaetoceros sp. Na11C3	Na11C3
	Na43A1
Chaetoceros sp. Na26B1	Na26B1
C. tenuissimus	GB2a
	Na26A1
	Na44A1

Table 1. List of strains utilised for single-strain HTS. GB2a is a strain from the Gulf of Naples. *Ch* Chile, *Na* Naples, *Ro* Roscoff.

patible with the GeneStudio S5 Ion Torrent (Life Technologies, Carlsbad, California). The obtained fragment contained all the information required for sequencing and differentiation (barcoding) of samples. A detailed description of the procedure is provided in Supplementary File S1.

Data pre-processing for single-strain HTS.

From raw fastq data, adapters and primers were removed with cutadapt[44], allowing a maximum of three mismatches. All reads with a length < 350 bp and quality score < 20 were discarded. Because data obtained with Ion Torrent technology are known to be sensitive to a high indel error rate in homopolymeric regions[45], we corrected indel errors using ICC v2.0.1[46].

Data analysis.

Data were analysed by means of abundance plots, analysis of similarity and phylogenetic haplotype networks. For computational and graphical reasons, we only considered for our analyses, when available, the first most abundant 50 haplotypes for both environmental and single-strain samples. As first data exploration, we plotted abundance patterns of the dominant haplotype over the minor haplotypes of validated species inferred from the environmental samples. Plots were made in R[47] using the packages *ggplot2*[48], *gridExtra*[49] and *scales*[50]. Furthermore, to ascertain if the aforementioned haplotype abundance patterns in both strains and environmental samples fitted existing distribution models, we explored the Rank Abundance Distribution (RAD) models using the radfit function in *vegan*[51]. This function fits the predictions of some of the most popular species abundance models (Broken stick Null model, Preemption model, Lognormal, Zipf and Zipf-Mandelbrot) to empirical data using maximum likelihood estimation.

As second step, we explored the pattern of similarity among the validated environmental haplotypes of *Chaetoceros* species, the reference barcodes and monoclonal single-strain data for each species using BLAST[39]. In particular, we assessed if: (i) the most abundant haplotype in each single strain matched the reference barcode of that strain obtained with Sanger sequencing and with the dominant environmental haplotype; (ii) the minor haplotypes in the strains were also found in the environmental samples. Furthermore, in order to assess the efficiency of homogenisation we calculated, for each strain, the percentage of reads clustering with the dominant haplotype. Clustering was performed at the threshold of 99% of similarity based on the findings of[52] using *vsearch* in mothur[43].

Finally, we inferred haplotype networks for each selected *Chaetoceros* species from temporal environmental data for a graphical visualisation of concerted evolution. If this phenomenon was occurring in our target species, we expected to see a major node (the dominant haplotype) surrounded by smaller ones whose temporal pattern (colour pattern in the nodes) was consistent. Networks were inferred using the TCS method[53] implemented in PopART v1.7[54]. Only metabarcodes with abundance ≥ 2 were used to reduce the number of sequences to be processed for network inference. Furthermore, for *C. costatus* and *C. tenuissimus*, we further reduced the number of haplotypes analysed considering only the ones with abundance ≥ 10 and ≥ 50 respectively, in order to obtain a clearer graphical visualisation of networks. Metabarcode dates were pooled together in months, and a different colour was assigned to each of them. To test if the temporal distribution of dominant and minor haplotypes in the networks was the same, we inferred the Kolmogorov–Smirnov test using PAST v3.24[55]. For the test we selected in each species' network, whenever possible, a few peripheral nodes with a distribution of reads over the months comparable (a colour pattern similar) to that of the dominant haplotype (nodes marked with *) as well as peripheral nodes with a markedly different distribution (nodes marked with #). We used 1000 Monte Carlo permutations for assessing the statistical significance of p values.

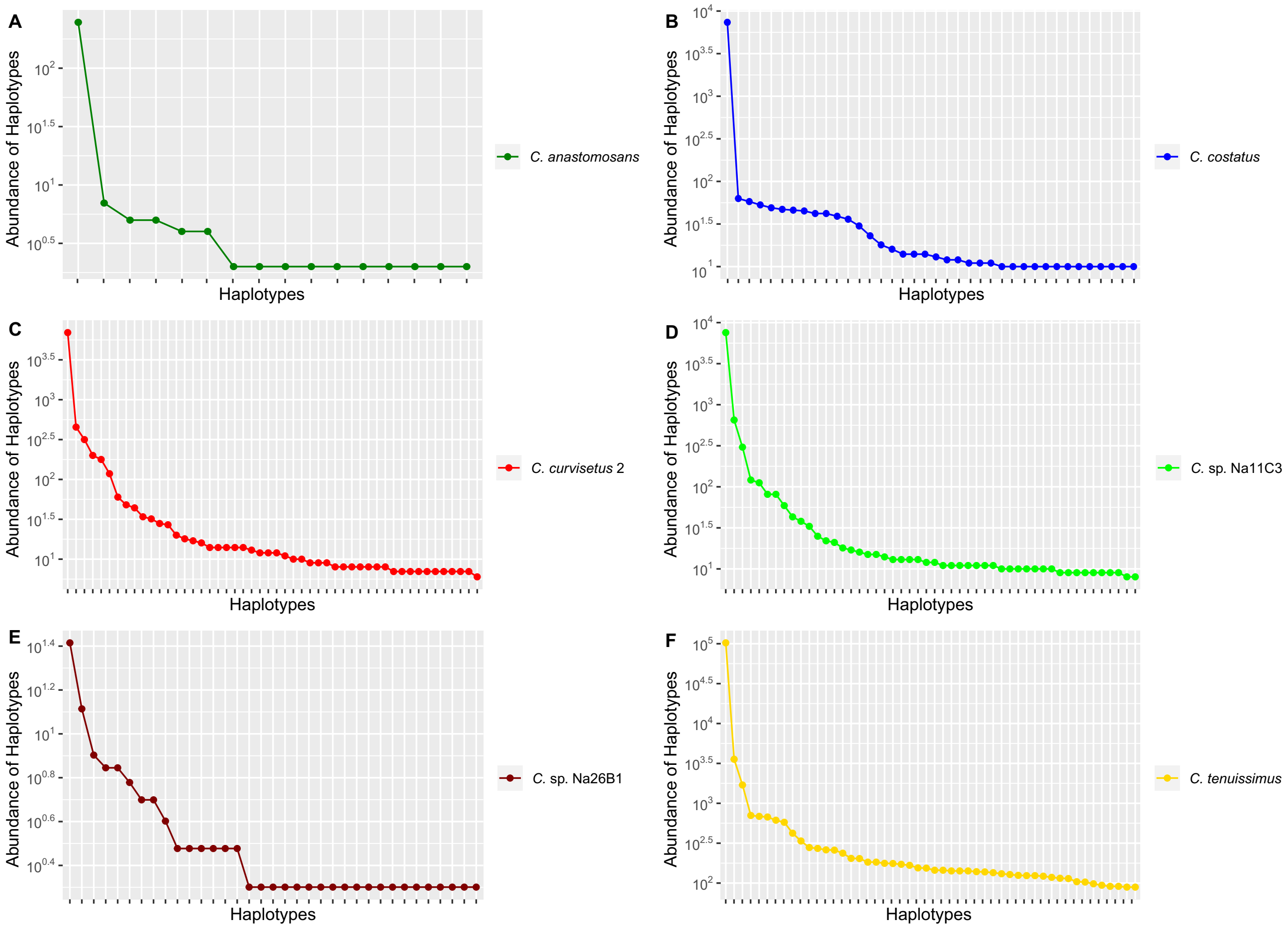

Figure 1. Abundance plots for each *Chaetoceros* species from validated environmental sequences. (**A**) *C. anastomosans*; (**B**) *C. costatus*; (**C**) *C. curvisetus* 2; (**D**) *Chaetoceros* sp. Na11C3; (**E**) *Chaetoceros* sp. Na26B1; (**F**) *C. tenuissimus*. Only the first 50 most abundant haplotypes, when available, were plotted in decreasing order of read number per haplotype. Data were from the temporal metabarcoding dataset "MareChiara" (January 2011 to December 2013).

Results

General characteristics of the datasets.

The number of reads and haplotypes (total number and sequences utilised for network analyses) for each species from the temporal environmental dataset after the validation procedure described in "Methods" section is provided in Supplementary Table S2. Briefly, the number of total validated reads ranged from 366 (90 haplotypes) in *C. anastomosans* to 139,185 (2585 haplotypes) in *C. tenuissimus*. The number of haplotypes utilised for network inference ranged from 15 (*C. anastomosans*) to 527 (*Chaetoceros* sp. Na11C3).

For single strain HTS, the number of reads ranged from 32,112 (*C. curvisetus* 2 Na1C1) to 516,766 (*Chaetoceros* sp. Na11C3) and, after pre-processing, from 19,185 (*C. curvisetus* 2 Na1C1) to 94,449 (*Chaetoceros* sp. Na11C3). The number of haplotypes used for following analyses ranged from a minimum of 2002 (*C. curvisetus* 2 strain Na1C1) to a maximum of 4696 (*C. costatus* strain Na32B1) (Supplementary Table S3). Raw data relative to single strain HTS can be found in NCBI Sequence Read Archive (SRA) at the accession numbers SAMN15700870–SAMN15700885.

Abundance plots from environmental metabarcoding and single strain HTS.

The plotting of the 50 most abundant haplotypes (Supplementary Table S4) from environmental metabarcoding data versus their abundance in each species (Fig. 1) showed a characteristic pattern. Indeed, in each species analysed, of all the haplotypes attributed to a particular species (environmental samples) there was one (the "dominant haplotype") that was far more abundant of all the others, of at least one order of magnitude (Fig. 1). All the other copies occurred in the environment at lower abundance. Patterns of abundance distribution in the HTS of single strains showed the same trend observed in the matabarcoding data of environmental samples (Fig. 2). Indeed, in each strain there was the same steep decrease in abundance from the dominant haplotype to the most abundant minor haplotype and then the more or less linear decrease along the minor haplotypes when scaled logarithmically. Furthermore, within the same species, the distribution of the 50 most abundant haplotypes among strains was congruent (Fig. 2) and most of minor haplotypes, despite not in the same ranking order, were identical

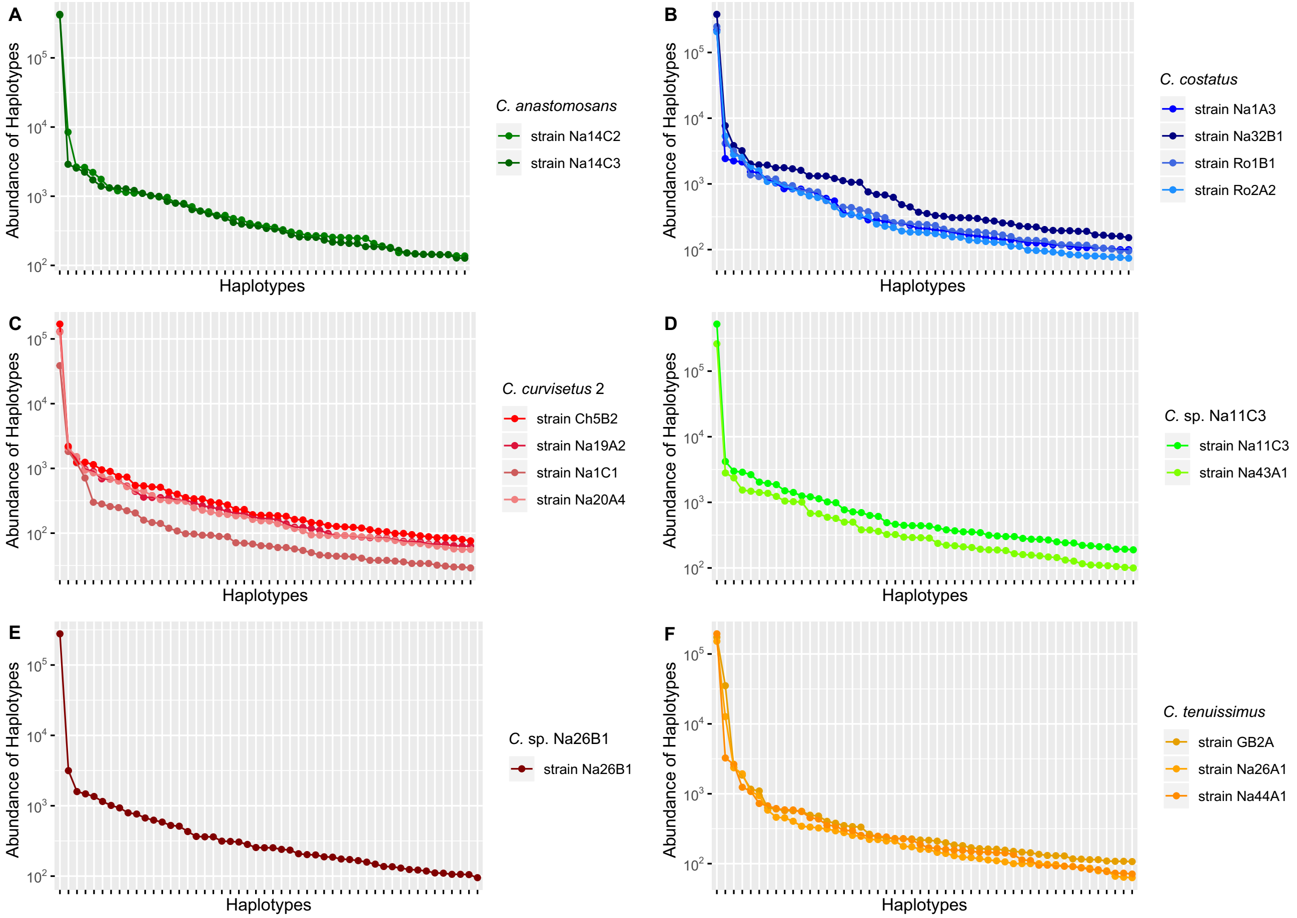

Figure 2. Abundance plots for each strain analysed in different *Chaetoceros* species. (**A**) *C. anastomosans*; (**B**) *C. costatus*; (**C**) *C. curvisetus* 2; (**D**) *Chaetoceros* sp. Na11C3; (**E**) *Chaetoceros* sp. Na26B1; (**F**) *C. tenuissimus*. Data are from single strain high throughput sequencing. Only the first 50 most abundant haplotypes were plotted.

		% similarity with haplotypes found in strains			
Species	**N environmental haplotypes**	**100**	**99.74–99.73**	**99.48–99.47**	**99.21–99.20**
C. anastomosans	14	42.9	92.9	100	–
C. costatus	38	73.7	100	–	–
C. curvisetus 2	369	53.6	95.1	100	–
C. sp. Na11C3	527	56.9	95.2	99.4	100
C. sp. Na26B1	59	45.8	95	100	–
C. tenuissimus	102	60.8	100	–	–

Table 2. Cumulative percentage of all environmental haplotypes found in the pooled single strains of each *Chaetoceros* species divided in classes of similarity.

across strains in this threshold (Supplementary Table S5). The comparison of models using Bayesian Information Criteria (BIC) highlighted in all the cases a preference for the Zipf model (Supplementary Figs. S1 and S2). Based on this model, the abundance of a haplotype was found to be inversely proportional to its rank and the expected abundance (a) of a haplotype at rank r is: $ar = N \cdot p1 \cdot r\gamma$, where N is number of reads, p1 is the fitted proportion of the most abundant haplotype, and γ is a decay coefficient. The Zipf-Mandelbrot model, a derivative of the Zipf model, was equally fitting our data (Supplementary Figs. S1 and S2).

Analyses of similarity. For each species, the reference barcode (obtained with Sanger sequencing) matched at 100% of similarity with the dominant haplotypes of environmental data and single strain HTS (Supplementary Table S6). In most of the species here examined, more than half of validated environmental minor haplotypes were also found in single strains HTS at 100% of similarity. Overall, a match between environmental and single strain haplotypes was found for each species within the similarity threshold of 99.20% (Table 2). The clustering analysis conducted in each strain to test the efficiency of homogenisation process showed that ca. 99% of the reads clustered with the dominant haplotype ≥ 99% similarity (Supplementary Table S7).

Phylogenetic networks from environmental samples. In all the species here analysed with a sufficient number of haplotypes (Fig. 3), the temporal pattern observed in the node containing the dominant haplotype corresponded with the temporal pattern of the other nodes containing haplotypes with lower abundance. This was particularly straightforward for *C. curvisetus* 2 (Fig. 3A), *Chaetoceros* sp. Na11C3 (Fig. 3B) and *C. tenuissimus* (Fig. 3C). These were also the species with the highest number of haplotypes utilised (369, 527 and 103 respectively). In *C. anastomosans* (Supplementary Fig. S3) this pattern is almost absent due to the low number of reads validated from the MareChiara dataset. In *C. tenuissimus* (Fig. 3C) the pattern of concerted evolution is particularly evident, which relates to the fact that it was the species with the highest number of haplotypes. Indeed, almost all the nodes around the central one containing the dominant haplotype showed a temporal pattern mimicking it. The inclusion of only the first 50 most abundant haplotypes in the analysis reduced the noise due to haplotypes at low read-abundances (e.g. less than 10) as observed in the networks of species with less abundant overall read numbers. As expected, the Kolmogorov–Smirnov test confirmed that the selected minor haplotypes with a read distribution over the months similar to that of the dominant haplotype (similar colour pattern) did not deviate significantly from that of the dominant haplotype ($p > 0.05$, Supplementary Table S8), whereas those with a strikingly different colour pattern deviated significantly ($p < 0.05$, Supplementary Table S8).

Discussion

Thanks to the experimental design presented in this study, we have demonstrated that the 18S rDNA region is under concerted evolution in the *Chaetoceros* species here analysed. Our results suggest that homogenisation is highly efficient at maintaining nearly identical 18S rDNA repeats. However, homogenisation remains incomplete as shown by the presence of minor haplotypes, which falls within the threshold of similarity of 99% in respect to the dominant haplotype. These results suggest that the 18S rDNA is evolving via concerted evolution rather than birth-and-death evolution[5]. Indeed, in studies reporting the occurrence of birth-and-death evolution, a greater genetic divergence among rDNA copies (around 10–15%) is observed[56–59], as consequence of the fact that some copies are randomly maintained in the genome for a long time while others are deleted[5]. Furthermore, we demonstrate that minor haplotypes found in the environmental samples are no technical artefacts because these same haplotypes are encountered not only in independently analysed samples from the same collection site but also in single strains within the threshold of 99.5% of similarity. We do not exclude that part of the minor variation is due to sequence error, but sequence error cannot explain the same sets of minor haplotypes all over the 48 independent samples and in the independently analysed strains.

The graphs obtained by plotting the first fifty most abundant haplotypes in temporal and single-strain HTS data also confirm the occurrence of concerted evolution. Indeed, both haplotypes from environmental metabarcoding and single strain HTS exhibit the same distribution pattern, with one haplotype that is far more abundant than all the others by at least one order of magnitude (the "dominant haplotype"), followed by other lowly abundant haplotypes ("minor haplotypes"). The fact that within the same species different strains share most of the minor haplotypes (though not in the same ranking order) is explained by the fact that concerted evolution requires not only the horizontal transfer of mutations among the repeats (homogenisation), but also the spread of mutations to all the individuals in the population (fixation)[7]. Furthermore, the fact that geographically distant strains share several minor haplotypes (e.g. *C. costatus* strains from Atlantic France and Mediterranean Naples, and *C. curvisetus* 2 strains from Central Chile and Mediterranean Naples) indicate that these haplotype variants were already present in the ancestral population from which these regional populations derived.

The pattern of haplotype abundance distribution here found best fits the Zipf- and Zipf-Mandelbrot (ZM) models among the RAD models tested. Evidence for these models exist in many biological systems[60–63]. Yet, the fitted distribution correctly describes only the first part of the empirical curves in both single strain and environmental data. Based on the abundance of the dominant haplotype, the Zipf-ZM models shapes a more rapid decay of the abundance of minor haplotypes given their rank. Several reasons are probably co-responsible for this discrepancy. Concerted evolution processes work towards homogenisation, but never fully achieve it. These models ideally fit an infinitely large dataset, whereas numbers of reads per haplotype are limited by our sample size, with many haplotypes on the right side of the curve with one read. A haplotype cannot include a fraction of a read, whereas the model can. In addition, we cannot exclude PCR and sequencing artefacts. Additional work is needed to ascertain if the Zipf-ZM models represent the best models for concerted evolution in other species across the Tree of Life. For this purpose, time-series metabarcoding data could be used to spot the signal of concerted evolution. However, these preliminary findings should be confirmed by other empirical data.

The presence of one dominant haplotype per species suggest that rDNA repeats in *Chaetoceros* are arranged in a single locus, but data on chromosomal distribution or number of nucleoli are needed to confirm this hypothesis. Regarding the copy number of rDNA cistron, several studies have demonstrated that it is highly variable: from 60 to 220 copies in fungi[64] and 39–19,300 in animals and 150–26,048 in plants[65]. Among protists, ciliates harbour the highest number of rDNA copies, between 3000 and 400,000[66], diatoms possess between 1057 and 12,812[67] and dinoflagellates between 200 and 1200[68]. In our study, assuming that each haplotype corresponds to a different rDNA unit, the number of repeats is around 3600 copies (min 2002; max 5055), within the values reported for diatoms. However, these estimations could overestimate the real number of copies due to the occurrence of PCR and sequencing errors. High variation among copies has been detected using a cloning and sequencing approach in fungi[69], dinoflagellates[70,71], and Foraminifera[28], as well as with genome sequencing in the plant genus *Asclepias*[72]. However, the biological relevance of having many rDNA haplotypes is largely unknown. Part

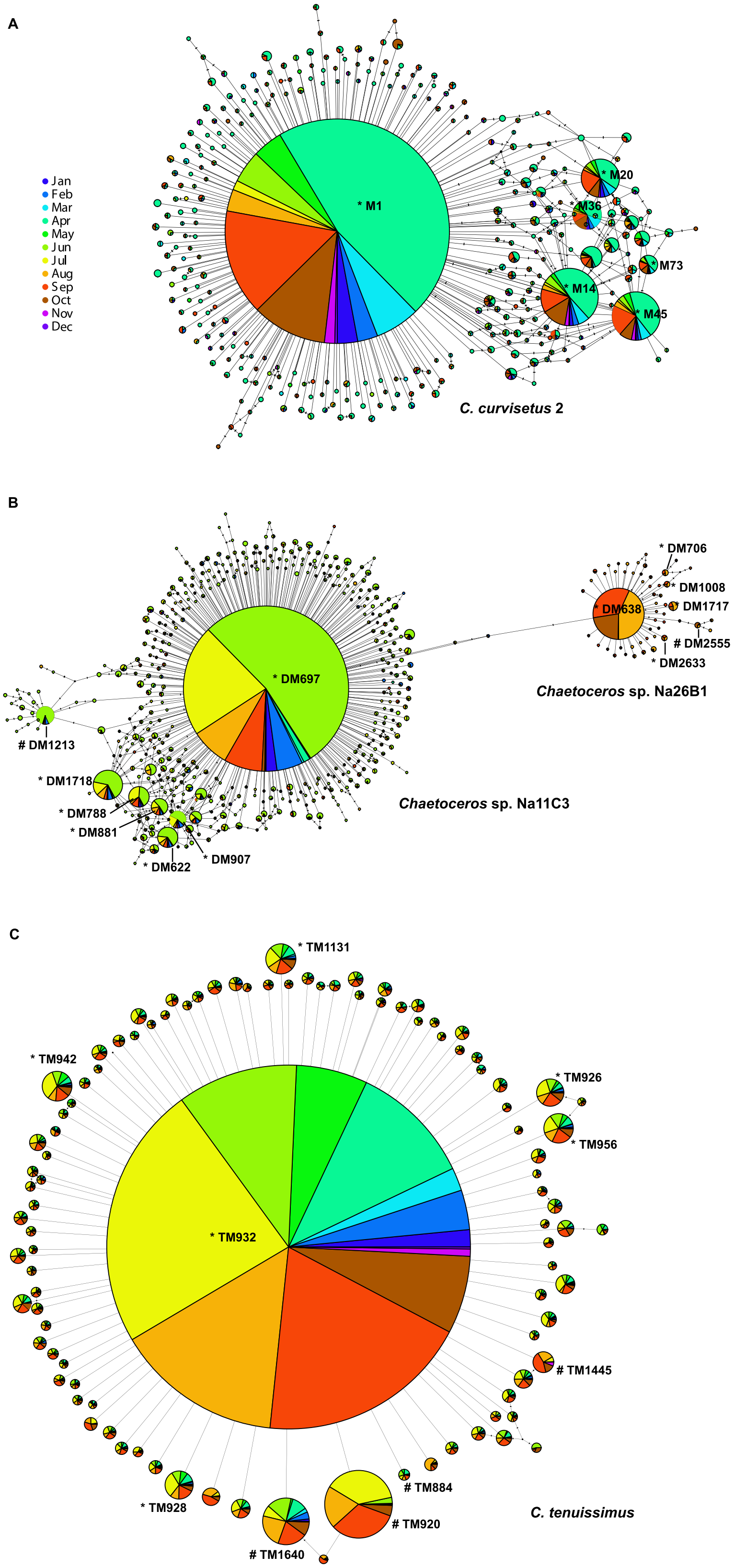

Figure 3. TCS haplotype networks inferred from environmental metabarcoding data. (**A**) *C. curvisetus* 2; (**B**) *Chaetoceros* sp. Na11C3 and *C.* sp. Na26B1; (**C**) *C. tenuissimus*. Haplotypes are grouped per month across 2011 and 2013 and partitioned as follows: 369 haplotypes (abundance ≥ 2) for *C. curvisetus* 2, 527 haplotypes for *C.* sp. Na11C3 and 59 for *C.* sp. Na26B1 (abundance ≥ 2) and 103 haplotypes (abundance ≥ 50) for *C. tenuissimus*.

of such variation could simply be due to imperfection of the homogenisation mechanism. Another explanation, complementary to the former, is that there could be a selective advantage in possessing all these different copies. Indeed, in bacteria it has been shown that the number of copies of the small rDNA gene correlates with the rate at which phylogenetically diverse bacteria respond to resource availability, with a high copy number leading to rapid colony formation[73]. In eukaryotes, the copy number of rDNA genes is unstable[74] and its stabilisation extends lifespan in yeast[75]. Always in yeast, it has been demonstrated recently that DNA replication stress induces a reduction in rDNA copy number in yeast[76]. The possible role of rDNA heterogeneity in protists is yet to be unveiled.

The results of analyses of similarity reveal that the dominant haplotypes of all the monoclonal strains analysed are identical within the same species, as well as to Sanger reference sequences and to the dominant metabarcode environmental haplotype of that species. Moreover, the minor haplotypes found in the environment data also occur in the monoclonal strains (intragenomic variation), and within the 99.5% of similarity in most of the cases. All these results provide empirical evidence for concerted evolution in these diatom species and, for the first time, by using single strain high-throughput sequencing and a metabarcoding time-series dataset together with a Sanger reference library.

The availability and the quality of Sanger reference barcode sequences is particularly important in the current "big biodiversity data" era, in which hundreds of millions of sequences can be generated during a single high-throughput sequencing run and the ability to check individual sequences is severely limited[77]. In the case of a multigene family subjected to effective concerted evolution, all the copies within a genome are expected to be identical and represented by the Sanger sequenced reference sequence. However, intragenomic sequence variation in multigene families is common in several eukaryotic lineages[19,69,78–80], and sometimes involves size variation[81,82]. In these cases, a Sanger sequence can be considered as a consensus of all the divergent copies of the amplified gene. In this "consensus sequence", most of the contribution is carried by the most abundant sequence, and therefore the Sanger sequence will read as the dominant haplotype.

Temporal metabarcoding datasets are classically used in ecological context to monitor biodiversity changes over time[38,83,84]. However, their potential goes beyond such purposes. In this study, we show that such datasets can be used also to track and test evolutionary or biological phenomena by the application of appropriate approaches. Indeed, through the inference of phylogenetic haplotype networks we obtain a graphical visualisation of concerted evolution in the *Chaetoceros* species here investigated. Each species network shows a star-shaped structure generated by intragenomic variation (lowly abundant divergent sequences corresponding to minor haplotypes) surrounding the dominant haplotype. Most of these nodes mimic the same temporal trend (seasonality) of the node containing the dominant haplotype, confirming the same origin and in accordance with the expectations of homogenisation process. Nevertheless, part of minor haplotypes shows deviations from this shared pattern. Such deviations, identified with a statistical test, can be mainly ascribed to artefacts due to PCR errors or by-product of massive parallel sequencing, even if a biological significance cannot be excluded and could be further explored.

Another novelty of our study is the use of single strain HTS data instead of clone libraries to study intragenomic variation and, specifically, concerted evolution. For example, the V4 region in the 18S gene is the currently recommended barcoding region for protists[25], whilst the ITS region serves as such for fungi[26]. Some authors[85–87] have argued that the concerted evolution process, known to affect ribosomal genes, may not be sufficiently effective to ensure complete sequence homogeneity. Therefore, knowing the extent of infraspecific variation and modality of evolution of such regions is vital to barcoding studies. Studies targeting the ribosomal genes in different organisms revealed the occurrence of several different copies within each organism analysed and highlighted the potential risk for barcoding studies[29,88]. Indeed, one of the characteristics of a good DNA barcode is to have high interspecific divergence and low intraspecific variability[89]. Dakal et al.[88] argued that the presence of several haplotypes within an individual shortens the barcoding gap and should be taken into consideration in barcoding studies of yeasts. However, what is lacking in these studies is information about the abundance of these "alternative" rDNA copies. Pillet et al.[28] tried to predict the number of haplotypes in each specimen of *Elphidium macellum* (Foraminifera) correlating the number of clones screened with the number of haplotypes found. The authors argued that although some of less abundant haplotypes could be due to PCR artefacts, the high Spearman correlation coefficient suggested that the real number of haplotypes in each individual was underestimated[28]. In this study, we demonstrate that within each strain of several *Chaetoceros* species occur thousands of 18S haplotypes, one of which is far more abundant that all the others (the "dominant" haplotype). Because of such huge differences in abundance, the probability that a "minor" haplotype is sequenced with Sanger chemistry is almost null. In turn, this means that there is no risk associated to the use of the rDNA cistron as target gene in classical DNA barcoding studies. However, in metabarcoding studies these minor haplotypes (intragenomic variation) can create a false rare diversity and therefore produce artefacts in diversity assessments[90]. This study also confirms, through the use of single strain HTS, the finding of[37] from environmental samples, i.e. that the most abundant haplotype that is recovered for each species corresponds to the sequence that would be obtained by Sanger sequencing. Therefore, in case of a taxon for which a reference sequence is not available yet, the dominant haplotype retrieved from a metabarcoding dataset can be considered as such, and subsequently validated using Sanger sequencing when the specimen has been sampled.

In conclusion, in this study we report the occurrence of concerted evolution in several *Chaetoceros* species through a specific experimental design based on plots of haplotype distribution, analyses of sequence similarity and evolutionary networks using Sanger reference sequences, environmental time-series metabarcoding data and single strain HTS. This approach is novel with respect to the classical one, in which concerted evolution is typically inferred indirectly from phylogenetic inferences[21,32,91,92]. We also show a novel use of metabarcoding and HTS data that goes beyond the traditional ecological applications. On the one hand, we confirm that the dominant haplotype perfectly matches with the Sanger reference sequence, validating the use of the metabarcoding technique for ecological studies. On the other hand, we highlight that the high number of sequences occurring at low abundances (minor haplotypes) inflate the diversity assessments, but they are intragenomic variation occurring in the strains. In this study, we show that at 99% of similarity, all infraspecific variability is collapsed together with the dominant haplotype. This is true for *Chaetoceros*, but the validity across other genera is to be tested yet. Finally, our study is also a first attempt to fit the biological phenomenon of concerted evolution to the most popular species abundance models. A possible course of action for future research could be to compare the results obtained in this study in *Chaetoceros* with other diatom and protist species, in order to understand the evolution of such gene region as well as the applicability of metabarcoding and high throughput sequencing in ecological and evolutionary studies in other marine organisms.

Data availability

Fastq files relative to single strain HTS are available on the Sequence Read Archive (SRA) database at the accession numbers SAMN15700870–SAMN15700885. Input data for phylogenetic haplotype networks are available on fighshare at https://doi.org/10.6084/m9.figshare.13299191. All other information is provided in the Supplementary Information file attached to this manuscript.

Received: 13 October 2020; Accepted: 28 December 2020
Published online: 12 January 2021

References

1. Britten, R. J. & Waring, M. Renaturation of the DNA of higher organisms. *Carnegie Inst. Washingt. Yearb.* **64**, 316–333 (1965).
2. Britten, R. J. & Kohne, D. E. Repeated sequences in DNA. *Science* **161**, 529–540 (1968).
3. Brown, D. D., Wensink, P. C. & Jordan, E. A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleri*: The evolution of tandem genes. *J. Mol. Biol.* **63**, 57–73 (1972).
4. ElderJohn, F. & Turner, B. J. Concerted evolution of repetitive DNA sequences in eukaryotes. *Q. Rev. Biol.* **70**, 297–320 (1995).
5. Nei, M. & Rooney, A. P. Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* **39**, 121–152 (2005).
6. Zimmer, E. A., Martin, S. L., Beverley, S. M., Kan, Y. W. & Wilson, A. C. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proc. Natl. Acad. Sci. U.S.A.* **77**, 2158–2162 (1980).
7. Graur, D. & Li, W. H. *Fundamentals of Molecular Evolution* (Sinauer Associates, Sunderland, 1991).
8. Coen, E., Strachan, T. & Dover, G. Dynamics of concerted evolution of ribosomal DNA and histone gene families in the melanogaster species subgroup of *Drosophila*. *J. Mol. Biol.* **158**, 17–35 (1982).
9. Ganley, A. R. D. & Kobayashi, T. Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Res.* **17**, 184–191 (2007).
10. Long, E. O. & Dawid, I. B. Repeated genes in eukaryotes. *Annu. Rev. Biochem.* **49**, 727–764 (1980).
11. Liebhaber, S. A., Goossens, M. & Kan, Y. W. Homology and concerted evolution at the α1 and α2 loci of human α-globin. *Nature* **290**, 26–29 (1981).
12. Gojobori, T. & Nei, M. Concerted evolution of the immunoglobulin VH gene family. *Mol. Biol. Evol.* **1**, 195–212 (1984).
13. Lindegren, C. C. Gene conversion in *Saccharomyces*. *J. Genet.* **51**, 625–637 (1953).
14. Holliday, R. A mechanism for gene conversion in fungi. *Genet. Res.* **5**, 282–304 (1964).
15. Charlesworth, B., Langley, C. H. & Stephan, W. The evolution of restricted recombination and the accumulation of repeated DNA sequences. *Genetics* **112**, 947–962 (1986).
16. Nikolaidis, N. & Nei, M. Concerted and nonconcerted evolution of the Hsp70 gene superfamily in two sibling species of nematodes. *Mol. Biol. Evol.* **21**, 498–505 (2004).
17. Andrea, L., Marini, M. & Mantovani, B. Non-concerted evolution of the RET76 satellite DNA family in *Reticulitermes* taxa (Insecta, Isoptera). *Genetica* **128**, 123–132 (2006).
18. Li, Y., Jiao, L. & Yao, Y.-J. Non-concerted ITS evolution in fungi, as revealed from the important medicinal fungus *Ophiocordyceps sinensis*. *Mol. Phylogenet. Evol.* **68**, 373–379 (2013).
19. Alverson, A. J. & Kolnick, L. Intragenomic nucleotide polymorphism among small subunit (18S) rDNA paralogs in the diatom genus *Skeletonema* (Bacillariophyta). *J. Phycol.* **41**, 1248–1257 (2005).
20. Harpke, D. & Peterson, A. Non-concerted ITS evolution in *Mammillaria* (Cactaceae). *Mol. Phylogenet. Evol.* **41**, 579–593 (2006).
21. Zheng, X., Cai, D., Yao, L. & Teng, Y. Non-concerted ITS evolution, early origin and phylogenetic utility of ITS pseudogenes in *Pyrus*. *Mol. Phylogenet. Evol.* **48**, 892–903 (2008).
22. Xiao, L.-Q., Möller, M. & Zhu, H. High nrDNA ITS polymorphism in the ancient extant seed plant *Cycas*: Incomplete concerted evolution and the origin of pseudogenes. *Mol. Phylogenet. Evol.* **55**, 168–177 (2010).
23. Vilnet, A., Konstantinova, N. & Troitsky, A. Molecular phylogenetic data on reticulate evolution in the genus *Barbilophozia* Löske (Anastrophyllaceae, Marchantiophyta) and evidence of non-concerted evolution of rDNA in *Barbilophozia rubescens* allopolyploid. *Phytotaxa* **49**, 6–22 (2012).
24. Xu, B., Zeng, X.-M., Gao, X.-F., Jin, D.-P. & Zhang, L.-B. ITS non-concerted evolution and rampant hybridization in the legume genus *Lespedeza* (Fabaceae). *Sci. Rep.* **7**, 40057 (2017).
25. Pawlowski, J. *et al.* CBOL protist working group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biol.* **10**, e1001419–e1001419 (2012).
26. Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 6241–6246 (2012).
27. Stoeck, T., Przybos, E. & Dunthorn, M. The D1–D2 region of the large subunit ribosomal DNA as barcode for ciliates. *Mol. Ecol. Resour.* **14**, 458–468 (2013).
28. Pillet, L., Fontaine, D. & Pawlowski, J. Intra-genomic ribosomal RNA polymorphism and morphological variation in *Elphidium macellum* suggests inter-specific hybridization in foraminifera. *PLoS ONE* **7**, e32373–e32373 (2012).
29. Naidoo, K., Steenkamp, E. T., Coetzee, M. P. A., Wingfield, M. J. & Wingfield, B. D. Concerted evolution in the ribosomal RNA cistron. *PLoS ONE* **8**, e59355–e59355 (2013).
30. Vogler, A. P. & DeSalle, R. Evolution and phylogenetic information content of the ITS-1 region in the tiger beetle *Cicindela dorsalis*. *Mol. Biol. Evol.* **11**, 393–405 (1994).
31. Buckler, E. S., Ippolito, A. & Holtsford, T. P. The evolution of ribosomal DNA divergent paralogues and phylogenetic implications. *Genetics* **145**, 821–832 (1997).
32. Gong, L., Shi, W., Yang, M., Si, L. & Kong, X. Non-concerted evolution in ribosomal ITS2 sequence in *Cynoglossus zanzibarensis* (Pleuronectiformes: Cynoglossidae). *Biochem. Syst. Ecol.* **66**, 181–187 (2016).
33. Kuhn, G. C. S., Küttler, H., Moreira-Filho, O. & Heslop-Harrison, J. S. The 1.688 repetitive DNA of Drosophila: Concerted evolution at different genomic scales and association with genes. *Mol. Biol. Evol.* **29**, 7–11 (2011).
34. Bueno, D., Palacios-Gimenez, O. M., Martí, D. A., Mariguela, T. C. & Cabral-de-Mello, D. C. The 5S rDNA in two *Abracris* grasshoppers (Ommatolampidinae: Acrididae): Molecular and chromosomal organization. *Mol. Genet. Genomics* **291**, 1607–1613 (2016).

35. Matyášek, R. *et al.* Next generation sequencing analysis reveals a relationship between rDNA unit diversity and locus number in *Nicotiana* diploids. *BMC Genomics* **13**, 722 (2012).
36. Belyayev, A. *et al.* Natural history of a satellite DNA family: From the ancestral genome component to species-specific sequences, concerted and non-concerted evolution. *Int. J. Mol. Sci.* **20**, 1201 (2019).
37. Gaonkar, C. C. *et al.* Species detection and delineation in the marine planktonic diatoms *Chaetoceros* and *Bacteriastrum* through metabarcoding: Making biological sense of haplotype diversity. *Environ. Microbiol.* https://doi.org/10.1111/1462-2920.14984 (2020).
38. Piredda, R. *et al.* Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean Long Term Ecological Research site. *FEMS Microbiol. Ecol.* **93**, fiw200 (2016).
39. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
40. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20**, 1160–1166 (2019).
41. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490–e9490 (2010).
42. Han, M. V. & Zmasek, C. M. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinform.* **10**, 356 (2009).
43. Schloss, P. D. *et al.* Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).
44. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).
45. Laehnemann, D., Borkhardt, A. & McHardy, A. C. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Brief. Bioinform.* **17**, 154–179 (2016).
46. Deng, W. *et al.* Indel and Carryforward Correction (ICC): A new analysis approach for processing 454 pyrosequencing data. *Bioinformatics* **29**, 2402–2409 (2013).
47. R Core Team. R: A language and environment for statistical computing (2019).
48. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, Berlin, 2016).
49. Auguie, B. gridExtra: Miscellaneous Functions for 'Grid' Graphics (2017).
50. Wickham, H. & Seidel, D. scales: Scale Functions for Visualization (2019).
51. Oksanen, J., Blanchet, F., Friendly, M. vegan: Community Ecology Package. R package version 2.4-3 (2016).
52. De Luca, D., Kooistra, W. H. C. F., Sarno, D., Gaonkar, C. C. & Piredda, R. Global distribution and diversity of *Chaetoceros* (Bacillariophyta, Mediophyceae): Integration of classical and novel strategies. *PeerJ* **7**, e7410 (2019).
53. Clement, M., Posada, D. & Crandall, K. A. TCS: A computer program to estimate gene genealogies. *Mol. Ecol.* **9**, 1657–1659 (2000).
54. Leigh, J. W. & Bryant, D. popart: Full-feature software for haplotype network construction. *Methods Ecol. Evol.* **6**, 1110–1116 (2015).
55. Hammer, D. A. T., Ryan, P. D., Hammer, Ø. & Harper, D. A. T. Past: Paleontological statistics software package for education and data analysis. *Palaeontol. Electron.* **4**, 1–9 (2001).
56. Carranza, S., Giribet, G., Ribera, C. & Riutort, M. Evidence that two types of 18S rDNA coexist in the genome of *Dugesia* (Schmidtea) mediterranea (Platyhelminthes, Turbellaria, Tricladida). *Mol. Biol. Evol.* **13**, 824–832 (1996).
57. Muir, G., Fleming, C. C. & Schlötterer, C. Three divergent rDNA clusters predate the species divergence in *Quercuspetraea* (Matt.) Liebl. and *Quercusrobur* L. *Mol. Biol. Evol.* **18**, 112–119 (2001).
58. Rooney, A. P. Mechanisms underlying the evolution and maintenance of functionally heterogeneous 18S rRNA genes in apicomplexans. *Mol. Biol. Evol.* **21**, 1704–1711 (2004).
59. Sipiczki, M., Horvath, E. & Pfliegler, W. P. Birth-and-death evolution and reticulation of ITS segments of *Metschnikowia andauensis* and *Metschnikowia fructicola* rDNA repeats. *Front. Microbiol.* **9**, 1193 (2018).
60. Li, W. & Yang, Y. Zipf's law in importance of genes for cancer classification using microarray data. *J. Theor. Biol.* **219**, 539–551 (2002).
61. Papp, L. & Izsák, J. Diversity and abundance relationships in a fly collection from a salt lake in central Hungary. *Commun. Ecol.* **9**, 99–105 (2008).
62. Spatharis, S. & Tsirtsis, G. Zipf-Mandelbrot model behavior in marine eutrophication: Two way fitting on field and simulated phytoplankton assemblages. *Hydrobiologia* **714**, 191–199 (2013).
63. Aitchison, L., Corradi, N. & Latham, P. E. Zipf's law arises naturally when there are underlying, unobserved variables. *PLoS Comput. Biol.* **12**, e1005110 (2016).
64. Simon, D., Moline, J., Helms, G., Friedl, T. & Bhattacharya, D. Divergent histories of rDNA group I introns in the Lichen family Physciaceae. *J. Mol. Evol.* **60**, 434–446 (2005).
65. Prokopowich, C. D., Gregory, T. R. & Crease, T. J. The correlation between rDNA copy number and genome size in eukaryotes. *Genome* **46**, 48–50 (2003).
66. Gong, J., Dong, J., Liu, X. & Massana, R. Extremely high copy numbers and polymorphisms of the rDNA operon estimated from single cell analysis of oligotrich and peritrich ciliates. *Protist* **164**, 369–379 (2013).
67. Godhe, A. *et al.* Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Appl. Environ. Microbiol.* **74**, 7174–7182 (2008).
68. Galluzzi, L. *et al.* Development of a real-time PCR assay for rapid detection and quantification of *Alexandrium minutum* (a Dinoflagellate). *Appl. Environ. Microbiol.* **70**, 1199–1206 (2004).
69. Simon, U. K. & Weiss, M. Intragenomic variation of fungal ribosomal genes is higher than previously thought. *Mol. Biol. Evol.* **25**, 2251–2254 (2008).
70. Gribble, K. E. & Anderson, D. M. High intraindividual, intraspecific, and interspecific variability in large-subunit ribosomal DNA in the heterotrophic dinoflagellates *Protoperidinium*, *Diplopsalis*, and *Preperidinium* (Dinophyceae). *Phycologia* **46**, 315–324 (2007).
71. Miranda, L. N., Zhuang, Y., Zhang, H. & Lin, S. Phylogenetic analysis guided by intragenomic SSU rDNA polymorphism refines classification of "*Alexandrium tamarense*" species complex. *Harmful Algae* **16**, 35–48 (2012).
72. Weitemier, K., Straub, S. C. K., Fishbein, M. & Liston, A. Intragenomic polymorphisms among high-copy loci: A genus-wide study of nuclear ribosomal DNA in *Asclepias* (Apocynaceae). *PeerJ* **3**, e718–e718 (2015).
73. Klappenbach, J. A., Dunbar, J. M. & Schmidt, T. M. rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.* **66**, 1328–1333 (2000).
74. Ganley, A. R. D. & Kobayashi, T. Ribosomal DNA and cellular senescence: New evidence supporting the connection between rDNA and aging. *FEMS Yeast Res.* **14**, 49–59 (2014).
75. Howitz, K. T. *et al.* Small molecule activators of sirtuins extend *Saccharomyces cerevisiae* lifespan. *Nature* **425**, 191–196 (2003).
76. Salim, D. *et al.* DNA replication stress restricts ribosomal DNA copy number. *PLoS Genet.* **13**, e1007006–e1007006 (2017).
77. Weigand, H. *et al.* DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *Sci. Total Environ.* **678**, 499–524 (2019).
78. Wörheide, G., Nichols, S. A. & Goldberg, J. Intragenomic variation of the rDNA internal transcribed spacers in sponges (Phylum Porifera): Implications for phylogenetic studies. *Mol. Phylogenet. Evol.* **33**, 816–830 (2004).
79. Bik, H. M., Fournier, D., Sung, W., Bergeron, R. D. & Thomas, W. K. Intra-genomic variation in the ribosomal repeats of nematodes. *PLoS ONE* **8**, e78230 (2013).
80. Weber, A. A. T. & Pawlowski, J. Wide occurrence of SSU rDNA intragenomic polymorphism in foraminifera and its implications for molecular species identification. *Protist* **165**, 645–661 (2014).

81. Whang, I.-J., Jung, J., Park, J.-K., Min, G.-S. & Kim, W. Intragenomic length variation of the ribosomal DNA intergenic spacer in a malaria vector, *Anopheles sinensis*. *Mol. Cells* **14**(1), 158–216 (2002).
82. Fernández-Tajes, J. & Méndez, J. Two different size classes of 5S rDNA units coexisting in the same tandem array in the razor clam *Ensis macha*: Is this region suitable for phylogeographic studies?. *Biochem. Genet.* **47**, 775–788 (2009).
83. Guardiola, M. *et al.* Spatio-temporal monitoring of deep-sea communities using metabarcoding of sediment DNA and RNA. *PeerJ* **2016**, e2807 (2016).
84. Salonen, I. S., Chronopoulou, P. M., Leskinen, E. & Koho, K. A. Metabarcoding successfully tracks temporal changes in eukaryotic communities in coastal sediments. *FEMS Microbiol. Ecol.* **95**, 226 (2018).
85. Chase, M. W. *et al.* A proposal for a standardised protocol to barcode all land plants. *Taxon* **56**, 295–299 (2007).
86. Sonnenberg, R., Nolte, A. & Tautz, D. An evaluation of LSU rDNA D1–D2 sequences for their use in species identification. *Front. Zool.* **4**, 6 (2007).
87. Spooner, D. M. DNA barcoding will frequently fail in complicated groups: An example in wild potatoes. *Am. J. Bot.* **96**, 1177–1189 (2009).
88. Dakal, T. C., Giudici, P. & Solieri, L. Contrasting patterns of rDNA homogenization within the *Zygosaccharomyces rouxii* species complex. *PLoS ONE* **11**, e0160744–e0160744 (2016).
89. Kress, W. J. & Erickson, D. L. DNA barcodes: Genes, genomics, and bioinformatics. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 2761–2762 (2008).
90. Lindner, D. L. & Banik, M. T. Intragenomic variation in the ITS rDNA region obscures phylogenetic relationships and inflates estimates of operational taxonomic units in genus *Laetiporus*. *Mycologia* **103**, 731–740 (2011).
91. Popp, M. & Oxelman, B. Evolution of a RNA polymerase gene family in *Silene* (Caryophyllaceae)—incomplete concerted evolution and topological congruence among paralogues. *Syst. Biol.* **53**, 914–932 (2004).
92. Peng, Y.-Y. *et al.* The evolution pattern of rDNA ITS in Avena and phylogenetic relationship of the *Avena* species (Poaceae: Aveneae). *Hereditas* **147**, 183–204 (2010).

Acknowledgements

The authors acknowledge R. Pannone and E. Mauriello from the Molecular Sequencing and Analysis Center (CSAM) of the Stazione Zoologica Anton Dohrn for guidance and assistance during single strains HTS. W.H.C.F.K. acknowledges funding from H2020 RI Cluster project EMBRIC (GA-654008) and FP7 project ASSEMBLE (GA-227799). D.D.L. was supported by a PhD fellowship funded by the Stazione Zoologica Anton Dohrn of Naples (The Open University—Stazione Zoologica Anton Dohrn PhD Program).

Author contributions

D.D.L., W.H.C.F.K., D.S. and R.P. conceived and designed the study. E.B. provided technical support to the set-up of single strain metabarcoding. D.D.L. and R.P. collected the data. D.D.L., W.H.C.F.K. and R.P. analysed the data. D.D.L., W.H.C.F.K. and R.P. drafted the initial version of the manuscript and all authors contributed to later versions of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-020-80829-6.

Correspondence and requests for materials should be addressed to D.D.L. or R.P.