Reverse transcriptase genes are highly abundant and transcriptionally active in marine plankton assemblages

Magali Lescot, Pascal Hingamp, Kenji K Kojima, Emilie Villar, Sarah Romac, Alaguraj Veluchamy, Martine Boccara, Olivier Jaillon, Daniele Iudicone, Chris Bowler, et al.

OPEN

The ISME Journal (2015), 1–13

www.nature.com/ismej

ORIGINAL ARTICLE

Reverse transcriptase genes are highly abundant and transcriptionally active in marine plankton assemblages

Magali Lescot[1], Pascal Hingamp[1], Kenji K Kojima[2], Emilie Villar[1], Sarah Romac[3,4], Alaguraj Veluchamy[5,11], Martine Boccara[5], Olivier Jaillon[6,7,8], Daniele Iudicone[9], Chris Bowler[5], Patrick Wincker[6,7,8], Jean-Michel Claverie[1] and Hiroyuki Ogata[10]

[1]*Information Génomique et Structurale, UMR7256, CNRS, Aix-Marseille Université, Institut de Microbiologie de la Méditerranée (FR3479), Parc Scientifique de Luminy, Marseille, France;* [2]*Genetic Information Research Institute, Los Altos, CA, USA;* [3]*CNRS, UMR 7144, team EPEP, Station Biologique de Roscoff, Place Georges Teissier, Roscoff, France;* [4]*Sorbonne Universités, UPMC Univ Paris 06, Station Biologique de Roscoff, Place Georges Teissier, FR-Roscoff, France;* [5]*Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Paris, France;* [6]*CEA-Institut de Génomique, GENOSCOPE, Centre National de Séquençage, Evry Cedex, France;* [7]*Université d'Evry, Evry Cedex, France;* [8]*Centre National de la Recherche Scientifique (CNRS), Evry Cedex, France;* [9]*Laboratory of Ecology and Evolution of Plankton, Stazione Zoologica Anton Dohrn, Naples, Italy and* [10]*Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, Japan*

Genes encoding reverse transcriptases (RTs) are found in most eukaryotes, often as a component of retrotransposons, as well as in retroviruses and in prokaryotic retroelements. We investigated the abundance, classification and transcriptional status of RTs based on *Tara* Oceans marine metagenomes and metatranscriptomes encompassing a wide organism size range. Our analyses revealed that RTs predominate large-size fraction metagenomes (>5 μm), where they reached a maximum of 13.5% of the total gene abundance. Metagenomic RTs were widely distributed across the phylogeny of known RTs, but many belonged to previously uncharacterized clades. Metatranscriptomic RTs showed distinct abundance patterns across samples compared with metagenomic RTs. The relative abundances of viral and bacterial RTs among identified RT sequences were higher in metatranscriptomes than in metagenomes and these sequences were detected in all metatranscriptome size fractions. Overall, these observations suggest an active proliferation of various RT-assisted elements, which could be involved in genome evolution or adaptive processes of plankton assemblage.
The ISME Journal advance online publication, 27 November 2015; doi:10.1038/ismej.2015.192

Introduction

Transposable elements (TEs) have been found in virtually all eukaryotes and 80% of prokaryotes sequenced so far (Hua-Van *et al.*, 2011). They are usually considered as selfish DNA with the capacity to proliferate inside the genome (Doolittle and Sapienza, 1980; Orgel and Crick, 1980). Most organisms have no efficient way of eliminating these potentially deleterious genetic elements from their genomes, although different mechanisms that silence the activity of TEs are being increasingly revealed (Galagan *et al.*, 2003; Slotkin and Martienssen, 2007; Siomi *et al.*, 2011; Watanabe *et al.*, 2013). In spite of their parasitic nature, TEs also provide beneficial effects on the evolution of their hosts (Dunlap *et al.*, 2006; Slotkin and Martienssen, 2007; Casacuberta and Gonzalez, 2013; Gifford *et al.*, 2013; Riordan and Dupuy, 2013; Bennetzen and Wang, 2014). Characterizing the distribution and classification of TEs is thus important to better evaluate their role in shaping the evolution, structure and function of genomes across the whole tree of life.

TEs are traditionally split into two different classes: Class I TEs (retrotransposons) and Class II TEs (DNA transposons), differing in their mode of transposition (Wicker *et al.*, 2007). Retrotransposons

Correspondence: Dr H Ogata Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto JP-611-0011, Japan.
E-mail: ogata@kuicr.kyoto-u.ac.jp
[11]Current address: Biological and Environmental Sciences and Engineering Division, Center for Desert Agriculture, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.
Received 27 April 2015; revised 27 August 2015; accepted 22 September 2015

transpose through reverse-transcribed RNA intermediates. Retroviruses and endogenous retroviruses belong to Class I (Wicker *et al.*, 2007). DNA transposons transpose within genomes without RNA intermediates. Retrotransposons have a 'copy-and-paste' transposition mechanism, whereas a majority of DNA transposons use a 'cut-and-paste' mechanism, although several Class II elements are known as 'copy-and-paste' DNA transposons. Retrotransposons are found in eukaryotes only, whereas DNA transposons are found in both eukaryotes and prokaryotes (Wicker *et al.*, 2007).

Retrotransposons are often responsible for the marked expansion of genomes especially in higher eukaryotes (for example, 42% of the human genome (Lander *et al.*, 2001), 76% in maize (Schnable *et al.*, 2009), 67% in the chromosome 3B of bread wheat (Choulet *et al.*, 2014); see (Arkhipova *et al.*, 2003; Rho *et al.*, 2010; Piednoel *et al.*, 2013) for invertebrate genomes). A significantly large number of retrotransposons are observed in the red seaweed *Chondrus crispus* where they represent 55% of the genome (Collen *et al.*, 2013). Retrotransposons are also found in unicellular eukaryotes, such as diatoms (Armbrust *et al.*, 2004; Bowler *et al.*, 2008), green algae (Derelle *et al.*, 2006; Blanc *et al.*, 2010, 2012) and choanoflagellates (Carr *et al.*, 2008) as well as in other aquatic or terrestrial protists (Bringaud *et al.*, 2007; Khan *et al.*, 2007; Lorenzi *et al.*, 2008; Bulman *et al.*, 2011). Their frequency in the genome can vary between related species. For example, retrotransposons occupy 5.8% of the genome of the pennate diatom *Phaeodactylum tricornutum*, but only 1.1% of the genome of the centric diatom *Thalassiosira pseudonana* (Bowler *et al.*, 2008). The genomes of *Ostreococcus* species harbor retrotransposons (Derelle *et al.*, 2006; Palenik *et al.*, 2007), whereas no TEs have been identified in the closely related *Bathycoccus prasinos* (Moreau *et al.*, 2012).

Prokaryotic genomes also encode reverse transcriptases (RTs), although these RTs are less characterized than eukaryotic RTs. About 25% of sequenced bacterial genomes encode RTs (Simon and Zimmerly, 2008), which could be classified into 17 groups (Toro and Nisa-Martinez, 2014). The most prevalent groups of prokaryotic RTs are encoded in genetic elements such as group II introns, retrons and diversity-generating retroelements (DGRs).

In the current work, we explored recently available marine metagenomes and identified numerous genes encoding RT. We generated metagenomic data from samples collected at three sites in the Mediterranean Sea during the *Tara* Oceans project (Karsenti *et al.*, 2011; Hingamp *et al.*, 2013; Bork *et al.*, 2015). The Mediterranean Sea is an evaporative, oligotrophic to ultra-oligotrophic basin (mostly due to phosphate deficiency (Berland *et al.*, 1980; Krom *et al.*, 1991; Antoine *et al.*, 1995)). Despite its small size, it has a relatively complex pattern of plankton trophic regimes (D'Ortenzio and Ribera d'Alcala, 2009). Overall, a decreasing west–east gradient characterizes the surface chlorophyll pattern, further modulated by the presence of river runoffs (especially in the Adriatic Sea) and by the circulation, highly constrained by the orography of the surrounding regions (D'Ortenzio and Ribera d'Alcala, 2009). The planktonic community is generally dominated by small autotrophs, microheterotrophs and egg-carrying copepod species (see review by (Siokou-Frangou *et al.*, 2010)). The metagenomic data were derived from size-fractionated samples and corresponded to a wide organism size range (0.2 μm–2 mm), thus providing a comprehensive access to the genomic diversity of prokaryotes, eukaryotes and viruses residing in these water masses. Metatranscriptomic sequence data were generated for all samples (above 0.8 μm). In this study, we investigated the abundance, classification and taxonomic origin of RT sequences detected in these samples, as well as their transcriptional status.

Materials and methods

Samples and sequence data

Samples were collected at three *Tara* Oceans stations: (1) Station (St) TARA_007 (37°1′16′′N, 1°56′58′′E; 23 September 2009), situated in the Algerian basin close to Algiers, (2) St TARA_023 (39°50′9′′N, 17°24′17′′E; 16 November 2009) in the Adriatic Sea close to Dubrovnik and (3) St TARA_030 (33°55′5′′N, 32°53′40′′E; 15 December 2009) in the East Levantine basin south of Cyprus. At these three stations, size-fractionated plankton samples (five fractions: pico-plankton (0.2–1.6 μm), piconano-plankton (0.8–5 μm), nano-plankton (5–20 μm), micro-plankton (20–180 μm) and meso-plankton (180–2000 μm)) were collected at two depths (surface (SUR) and deep chlorophyll maximum (DCM)). Sampling protocols and environmental data associated with each sample are available in (Tara Oceans Consortium, Coordinators; Tara Oceans Expedition, Participants, 2014; Chaffron *et al.*, 2014a,b; Pesant *et al.*, 2015). Methods for DNA and RNA extraction, complementary DNA synthesis and sequencing library preparation are provided in Supplementary Methods and Results.

Twenty-nine DNA samples (five size fractions) were sequenced using Illumina HiSeq technology (one sample (TARA_030/DCM/180–2000 μm) did not yield enough DNA to be processed). For two samples (TARA_007/SUR and DCM/5–20 μm), DNA yield was low, therefore, DNA was whole genome amplified before sequencing. Sequence data from these two samples were used only in part of the present study (they were not included in quantitative analyses). The metagenome samples yielded 4714 million reads. For assembly, 3160 million reads were used to generate 4.2 Gbp of assembled sequences in five million metagenomic contigs (Supplementary Table S1), with the use of either SOAPdenovo v1.4/v1.5 (Luo *et al.*, 2012) or Velvet v1.0.15

(Zerbino and Birney, 2008). Only contigs $\geqslant 500$ bp were included in these assemblies. Open reading frames (ORFs) in the metagenomic contigs were identified by a combination of *de novo* gene-prediction methods (that is, metagene (Noguchi *et al.*, 2006) and SNAP (Korf, 2004)) (Supplementary Table S2).

Twenty-three RNA samples (four size fractions above 0.8 μm) were sequenced using Illumina HiSeq technology (one sample TARA_030/DCM/180–2000 μm did not yield enough RNA to be processed). A total of 1 776 979 581 reads from polyA mRNA samples yielded 8.0 Gbp of assembled sequences in 66 million metatranscriptomic contigs (Supplementary Table S1), with the use of either Velvet v1.2.07 or Oases v0.2.8 (Schulz *et al.*, 2012). Only contigs $\geqslant 500$ bp were included in these assemblies. We identified ORFs (≥30 amino acids (aa)) in metatranscriptomic contigs using EMBOSS/GETORF (Rice *et al.*, 2000).

More detailed descriptions of samples, sequence data and assembly processes are provided elsewhere (Jaillon *et al.*, submitted). All environmental sequences presented in this study are available at the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under the project identification number PRJEB402. Additional data are available from ftp://ftp.genome.jp/pub/db/community/tara/RT/.

Domain identification in metagenomes

Metagenomic ORFs were first masked using the SEG low-complexity sequence filter (Wootton, 1994). Using position-specific iterated basic local alignment search tool (Altschul *et al.*, 1990, 1997), we performed sequence similarity searches from all position-specific score matrices in the Conserved Domain Database (CDD) of the National Center for Biotechnology Information (NCBI) (Marchler-Bauer *et al.*, 2013) to the translated and filtered ORFs. Then the best CDD profile (E-value $< 10^{-5}$) was assigned to each ORF. To compile a list of retrotransposon- and retrovirus-related sequence domains recorded in CDD, the definition lines of CDD profiles were searched for relevant keywords (for example, 'retrotransposon', 'retrovirus', 'retroviral', 'RNase', 'integrase'). We manually examined the candidates and produced the final list containing 56 CDD profiles (Supplementary Table S3). These CDD profiles were used to identify retrotransposon- and retrovirus-related sequence domains in the metagenomic ORFs. It should be noted that, for the analysis of RT-containing ORFs, we used hidden Markov models (HMMs) instead of position-specific iterated basic local alignment search tool (see below).

Classification of the RTs

We built seven HMM profiles (HMMER/hmmbuild (Finn *et al.*, 2011)) based on multiple sequence alignments of diverse groups of RT sequences retrieved from the Gypsy database (http://gydb.org; (Llorens *et al.*, 2011)) as well as from (Gladyshev and Arkhipova, 2011). Searches with these seven profiles plus 51 HMM pre-compiled profiles from Gypsy database were carried out against the metagenomic/metatranscriptomic ORFs (using hmmsearch, E-value $< 10^{-5}$). Identified metagenomic sequences were compared by BLASTP (E-value $< 10^{-5}$) with the reference RT sequences (Gypsy database and (Gladyshev and Arkhipova, 2011)) and the best BLAST hit was used for RT classification (for example, LINEs, Gypsy, Copia and so on). The longest RT-like ORF from the *Tara* Oceans metagenomes (AHX23DCM1GGMM11BCE.ASY1CTG52) and its close homologs (cryptophyte *Dualen-1_GCr* and *Dualen-5_CCu*) that we identified in the Marine Microbial Eukaryote Transcriptome Sequencing Project (http://camera.crbs.ucsd.edu/mmetsp/, (Keeling *et al.*, 2014), as of May 2014) were submitted to Repbase (http://www.girinst.org/repbase/; (Jurka *et al.*, 2005)). Relative gene abundances in the metagenomic and metatranscriptomic data sets were estimated by their 'average coverage' defined by the cumulative sizes of reads mapped on the contigs divided by contig sizes (see Supplementary Methods and Results for detail).

Phylogenetic analysis

We aligned long metagenomic RT sequences (from 497 to 2543 aa) on the previously described reference HMM profiles using HMMER/hmmalign. We discarded alignment columns with $>80\%$ of gaps, and the resulting multiple sequence alignments used for phylogenetic analyses contained 462 positions (214 sequences) for the RT groups that included LINEs and Gypsy (Figure 1a), 257 positions (171 sequences) for Copia RTs (Figure 1b) and 315 positions (132 sequences) for BEL RTs (Figure 1c). The phylogenetic reconstructions were performed using the Phylogeny.fr server (Dereeper *et al.*, 2008) with PhyML 3.0 (Guindon *et al.*, 2010). The best fitting substitution model and rate variation parameters were selected using ProtTest 3 (Darriba *et al.*, 2011) according to the smallest Akaike Information Criterion: VT+I+G+F for the RT sequence set that include LINEs and Gypsy (Figure 1a), LG+I+G+F for Copia RTs (Figure 1b), RtREV+I+G+F for BEL RTs (Figure 1c). Bootstrap values were calculated with 100 bootstrap replicates. The resulting phylogenetic trees were edited using MEGA6 (Tamura *et al.*, 2013).

Criteria for environmental clades

To delineate 'environmental clades' that were only comprised of environmental sequences, we used the following three criteria: (i) the clade contained at least three environmental sequences, (ii) there was no known RT sequences (that is, RT sequences of known taxonomic origin from the public databases

such as NCBI non-redundant database (NR)) within the clade and (iii) the bootstrap branch support for the clade was ⩾80%. Prior to applying the second criterion, we first identified candidates for environmental clades fulfilling criteria (i) and (iii). We then performed BLASTP searches from the RT domains of the environmental sequences of the candidate clades against three databases: NR, our RT reference database (see above) and themselves. Most of the members of the initial candidate clades had BLASTP scores with other members of the same environmental clade higher than with any other known sequences, but several environmental sequences had best hits to known RT sequences. As the latter case may not satisfy the criterion (ii), these known sequences were added in our phylogenetic analyses. Only clades that satisfied the three criteria in the final phylogenetic trees were annotated as 'environmental clades'.

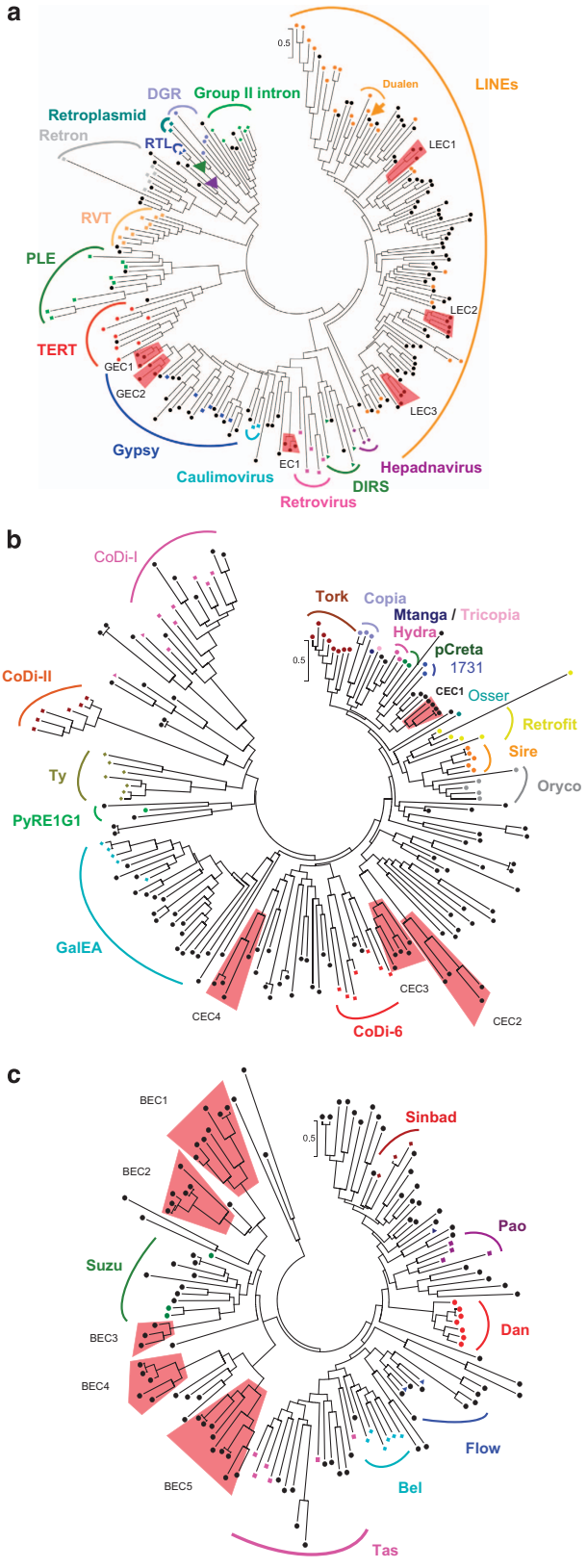

Figure 1 ML-trees of known and environmental RT sequences. (**a**) ML-tree of known RT sequences with 124 RT-like sequences identified in the metagenomic data. Black dots indicate RT-like sequences from the metagenomic data. EC1 stands for Environmental Clade 1, LEC for LINE Environmental Clades (LEC1-LEC3) and GEC for Gypsy Environmental Clades (GEC1, GEC2). LTR retrotransposons: Gypsy, Copia, BEL and retrovirus. Non-LTR retrotransposons: LINE (APE-type and REL-type). Distributions know reference RTs in prokaryotes (P) and eukaryotes (E) are as follows: LINE (E), Gypsy (E), Caulimovirus (E), Retrovirus (E), PLE (E), DIRS (E), RTL (E), RVT (E/P), TERT (E), group II intron (E/P), DGR (P), retron (P), retroplasmid (E), Hepadnavirus (E). ANB7SUR0CCII11BDE.ASY2CTG1927.ANO1.snap_1 is marked by a purple arrow, and AHX7DCM1GGMM11BCE.ASY1CTG361.ANO1.mga_1 by a green arrow. The sequence marked by an orange arrow is AHX23DCM1GGMM11BCE.ASY1CTG52.ANO1.mga_1. (**b**) ML-tree of known RT sequences with 100 Copia RT-like sequences identified in the metagenomic data. Black dots indicate RT-like sequences from the metagenomic data. CEC stands for Copia Environmental Clades (CEC1-CEC4). Copia belong to LTR retrotransposons. Taxonomic distributions of known reference Copia are as follows: CoDi (diatoms), GalEA (animals and red algae), PyRE1G1 (red algae), Hydra (Cnidarian), Tork (land plant), Mtanga and Tricopia (insecta), copia (insecta), pCreta (fungi), 1731 (diptera), Osser (plant), Retrofit (plant), Sire (plant), Oryco (plant), Ty (fungi). (**c**) ML-tree of known RT sequences with 100 longest BEL RT-like sequences identified in the metagenomic data. Black dots indicate RT-like sequences from the metagenomic data. BEC stands for BEL Environmental Clades (BEC1-BEC5). BEL belong to LTR retrotransposons. Tas (nematoda and cnidaria), Bel (insecta), Pao (insecta and vertebrata), Sinbad (invertebrata), Suzu (echinodermata and vertebrata). New environmental clades satisfying the three criteria defined in Materials and Methods are marked by a red-shaded area.

Taxonomic annotation

Metagenomic and metatranscriptomic RT-like ORFs were compared using BLASTP to reference databases (UniProt (release of May 2014) and Marine Microbial Eukaryote Transcriptome Sequencing Project). When the best hit reference sequence showed ⩾60% amino-acid sequence identity against a query sequence (E-value$<10^{-8}$), the taxonomic annotation of the reference sequence was transferred to the query sequence. This percent identity threshold was determined according to the distribution of sequence similarities among known Copia and Gypsy RT sequences to maximize the accuracy of taxonomic

annotation at the second level of the hierarchy of the NCBI taxonomic classification such as Stramenopile and Viridiplantae (Supplementary Figure S1).

Detrended correspondence analysis

To estimate the biotic (that is, organismal) and abiotic (that is, physicochemical) variables explaining patterns of RT distribution across samples, we used a detrended correspondence analysis (DCA) and the envfit function of the R package vegan (Oksanen *et al.*, 2007). Abiotic variables were available via PANGAEA (Tara Oceans Consortium, Coordinators; Tara Oceans Expedition, Participants, 2014; Chaffron *et al.*, 2014a,b) through the sample codes in Supplementary Table S4. Biotic composition of samples were derived from the 18S-V9 rDNA barcodes described in de Vargas *et al.* (2015), Supplementary Table S5). Vector fitting (envfit) with 1000 permutations was used to estimate the significance of correlations between the ecological variables and the ordination by DCA (P-values < 0.05).

Results

Characteristics of the sampling sites

The Mediterranean Sea shows a relatively complex pattern of plankton trophic regimes (D'Ortenzio and Ribera d'Alcala, 2009). Although not being representative of all the possible trophic regimes, the three sampling stations are related to quite different biomes. The first station, TARA_007, sampled the incoming (modified) Atlantic waters along the Algerian coast. These relatively fresh waters are separated from the deep layers by an intense pycnocline that prevents deep convection in winter, thus inhibiting the renewal of nutrients from the deep reservoir. Here, typical early fall conditions characterized the water column, with a shallow mixed layer (18 m) and mild temperatures (23.8 °C at surface). A shallow DCM (44 m) had a chlorophyll value of 0.5 mg Chl m^{-3}, that is, five times higher than the surface values. Despite the late fall sampling, TARA_023, in the more productive central Adriatic Sea, shows a very shallow mixed layer (9 m), possibly owing to the recent arrival of coastal waters from Croatia at the very surface. It has a surface temperature of 17.6 °C and a relatively deep DCM (54 m). Similarly sampled in late fall, TARA_030, at the center of the Cyprus gyre in the ultra-oligotrophic Eastern Basin, was characterized by a warmer surface temperature (20.4 °C) and a shallow mixed layer (41 m), overlying a DCM at 77 m. Both SUR and DCM chlorophylls were very low, as usual in this region (0.04 and 0.14 mg Chl m^{-3}, respectively).

RTs dominate the metagenomics gene content in size fractions above 5 μm

ORFs from 29 metagenomic assemblies were searched against NCBI/CDD. Out of 7 598 631 ORFs, 43% (3 236 369 ORFs) showed significant sequence similarity to the domain database. Of these, 18 780 ORFs were related to retrotransposons or retroviruses (Supplementary Table S3). Some of these CDD domains were ranked at the highest abundance levels. For instance, 9 of the 15 CDD profiles that recruited the largest numbers of metagenomic ORFs from the 180–2000 μm size fraction corresponded to retrotransposon/retrovirus-related domains (Table 1). The protein domains representing the RT of non-long terminal repeat (LTR) retrotransposons (cd01650) was ranked at the top (that is, 2782 ORFs). Other RT domains (cd01647, cd01644, pfam07727 and pfam00078) recruited from 222 to 1207 ORFs. Retrotransposon/retrovirus-related integrases (pfam00665), RNase H (pfam05380 and cd06222), and apurinic-like endonucleases (pfam03372) showed a large number of hits (from 186 to 988 ORFs). Other highly ranked domains unrelated to retrotransposons included ankyrin repeats, C2H2-type Zn-finger proteins, trypsin-like serine proteases, PIF1 helicases, homeobox transcription factor SIP1 and nidogen and related basement membrane proteins (Table 1). These are known to be encoded in multicopy genes. High abundances of retrotransposon/retrovirus-related domains were clearly observed in the three largest size fractions (that is, 180–2000 μm, 20–180 μm and 5–20 μm fractions) (Supplementary Figure S2 and Supplementary Table S6), with RTs of non-LTR retrotransposons (cd01650) and LTR retrotransposons including retroviruses (cd01647) being the most highly represented domains.

To classify and characterize the diversity of these retrotransposon/retrovirus-related sequence domains, we specifically focused on RT domains for the subsequent analyses. We identified 11 419 RT-like ORFs in the 29 metagenomes. The size of the RT-like metagenomic ORFs varied from 30 to 2543 aa (200 aa on average), whereas the length of typical known RTs is between 150 aa and 250 aa. The detected RT-like sequences were thus often incomplete, but many of them were long enough for phylogenetic analysis.

Phylogenetic trees indicated that the metagenomics-derived RTs were widely distributed over diverse RT groups (Figure 1 and Supplementary Figure S3). The tree in Figure 1a classified all RT-like sequences except Copia (Figure 1b) and BEL (Figure 1c) RTs. A majority of the metagenomic RTs were grouped with RTs from eukaryotes, with a high representation within the LINE and Gypsy groups. We identified three LINE clades that were represented only by environmental sequences (LINE Environmental Clades: LEC1 to 3), two Gypsy Environmental Clades (GEC1 and 2) and another environmental clade (EC1). Several sequences fell into groups of prokaryotic retroelements. For instance, one ORF (ANB7SUR0CCII11BDE.ASY2CTG1927.ANO1.snap_1, 0.2–1.6 μm fraction) marked in Figure 1a encoding an RT and DNA polymerase A resembles the Group D prokaryotic

Table 1 Highly represented CDD profiles in the ORF set from the 180–2000 µm size fraction metagenomes

Accession/name	*Description*	*Number of assigned ORFs*
cd01650/RT_nLTR_like[a]	Non-LTR retrotransposon and non-LTR retrovirus RT. This subfamily contains both non-LTR retrotransposons and non-LTR retrovirus RTs.	**2782**
cd01647/RT_LTR[a]	RTs from retrotransposons and retroviruses, which have LTRs in their DNA copies but not in their RNA template.	**1207**
pfam00665/rve[a]	Integrase core domain. Integrase mediates integration of a DNA copy of the viral genome into the host chromosome. Integrase is composed of three domains.	**988**
pfam03372/Exo_endo_phos[a]	Endonuclease/exonuclease/phosphatase family. This large family of proteins includes magnesium-dependent endonucleases and a large number of phosphatases involved in intracellular signaling.	**556**
cd01644/RT_pepA17[a]	RRTs in retrotransposons. This subfamily represents the RT domain of a multifunctional enzyme.	**551**
cd00204/ANK	Ankyrin repeats; ankyrin repeats mediate protein–protein interactions in very diverse families of proteins.	438
pfam05380/Peptidase_A17[a,b]	Pao retrotransposon peptidase. Corresponds to Merops family A17.	**431**
KOG2462/KOG2462	KOG2462, C2H2-type Zn-finger protein (Transcription).	342
cd00190/Tryp_SPc	Trypsin-like serine protease; many of these are synthesized as inactive precursor zymogens that are cleaved during limited proteolysis to generate their active forms.	321
pfam07727/RVT_2[a]	RT (RNA-dependent DNA polymerase). A RT gene is usually indicative of a mobile element such as a retrotransposon or retrovirus.	**260**
pfam05970/DUF889	PIF1 helicase. The PIF1 helicase inhibits telomerase activity and is cell cycle regulated.	249
pfam00078/RVT_1[a]	RT (RNA-dependent DNA polymerase).	**222**
KOG3623/KOG3623	KOG3623, Homeobox transcription factor SIP1 (Transcription).	211
cd06222/RNase H[a]	RNase H (RNase HI) is an endonuclease that cleaves the RNA strand of an RNA/DNA hybrid in a not sequence-specific manner.	**186**
KOG1214/KOG1214	KOG1214, nidogen and related basement membrane protein.	179

Abbreviations: CDD, Conserved Domain Database; LTR, long terminal repeat; ORF, open reading frame; RT, reverse transcriptase.
[a]CDD profiles representing proteins related to retrotransposons and retroviruses. For these profiles, the number of assigned ORFs are shown in bold letters.
[b]A recent report demonstrated that pfam05380 is actually an RNase H domain similar to cd06222 (RNase H) (Majorek *et al*., 2014).

retroelements in sequence and structure (Kojima and Kanehisa, 2008). Another sequence (AHX7DCM1 GGMM11BCE.ASY1CTG361.ANO1.mga_1, 0.8–5 µm fraction) was related to Abi (abortive bacteriophage infection) genes (Toro and Nisa-Martinez, 2014), without further precision about its origin. The metagenomic Copia RT sequences were distributed across the GalEA (marine arthropods) and three known groups of diatom Copia elements (CoDi-I (CoDi2.4, 3.1, 4.1, 4.3, 4.4, 4.5 and 7.1), CoDi-II (CoDi5.1 to 5.6) and CoDi6 (CoDi6.1 to 6.7) (Maumus *et al*., 2009)) (Figure 1b). Several RT-like sequences did not show close relatives in the reference Copia RT sequence set (Copia Environmental Clades: CEC1–4). Many of the metagenomic BEL RT sequences were grouped within Pao, Sinbad, Suzu, Tas and Bel retroelements (Figure 1c). Our phylogenetic analysis revealed the existence of five deeply branched and clearly delineated environmental clades for the BEL group (BEL Environmental Clade: BEC1–5).

Relative abundances of metagenomic RT sequences (based on average contig coverage) revealed that they were more abundant in the meso-, micro- and nano-plankton samples (5–2000 µm) than in the pico- and piconano-plankton samples (0.2–5 µm) (Figure 2). For three samples, the relative abundance of RT-like ORFs alone exceeded 8% of the abundance of all ORFs in the respective samples (that is, St TARA_023/DCM/180–2000 µm, 13.47%; St TARA_030/SUR/180–2000 µm, 9.02%; St TARA_030/DCM/5–20 µm, 8.73%).

Most metagenomics RT sequences originate from retrotransposons

We then classified metagenomic RT sequences (11 419 ORFs) (Figure 2). Consistent with our phylogenetic analyses for selected sequences, RTs from autonomous retrotransposons such as Gypsy, BEL and LINE-AP/LINE-REL were found to dominate in meso-, micro- and nano-plankton size fractions.

Taxonomic annotation suggested that RT-like sequences in large-size fractions (>5 µm) mainly originated from eukaryotic genomes (Supplementary Figure S4 and Supplementary Figure S5). The most represented groups were metazoans, followed by stramenopiles. In the piconano-plankton size fraction (0.8–5 µm), the taxonomic coverage was wider, with the detection of over 10 RT-like sequences in haptophytes, stramenopiles and cryptophytes than in the larger size fractions. Copia RTs were enriched in this piconano-plankton size fraction relative to other larger size fractions.

The longest ORF, AHX23DCM1GGMM11BCE. ASY1CTG52.ANO1.mga_1, (2543 aa; St TARA_023/ DCM/0.8–5 µm; Figure 1a) encompassed an apurinic-like endonuclease domain in addition to an RT domain and was found related to the cryptophyte LINEs, such as *Dualen-1_GCr* and *Dualen-5_CCu,* which we reconstructed from transcriptome sequences derived from Marine Microbial Eukaryote Transcriptome Sequencing Project (Keeling *et al*., 2014) (Supplementary Figure S6).

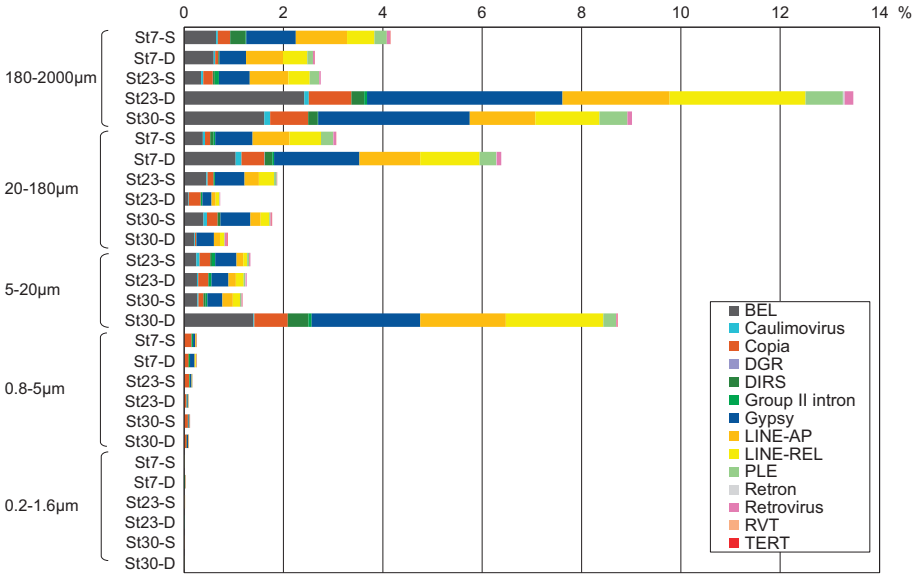

Figure 2 Relative RT gene abundance and classification of the RTs identified in the metagenomic data. In the station names, S and D denote, respectively, SUR and DCM depths.

Dualen-1_GCr from *Geminigera cryophila* encodes a 5287 aa protein including SET methyltransferase, C48 peptidase, apurinic-like endonuclease, reverse transcriptase, ribonuclease H, CCHC zinc finger and restriction-like endonuclease domains. This domain configuration is unique to the cryptophyte *Dualen* among other previously characterized *Dualen* elements (Kojima and Fujiwara, 2005). The main characteristic of the *Dualen* LINEs (also known as *RandI* (Kapitonov *et al.*, 2009)) is to encode both apurinic-like and restriction-like endonucleases (dual-endonucleases). The longest ORF with an RT-like domain thus likely represents a group of *Dualen* elements distributed among cryptophytes.

The relative abundance of bacterial-like environmental RT sequences increased with decreasing sample fraction size, with a high abundance of RTs of putative proteobacterial origin in line with the known overrepresentation of this taxon in the ocean. RT-like ORFs of predicted bacterial origin corresponded mainly to group II introns and retrons (Supplementary Figure S5).

Bacterial and viral RT transcripts were detected in all size fractions

Metatranscriptomic data were screened for the presence of transcriptionally active RT-like ORFs using HMM profiles. The search resulted in the identification of 4 258 RT-like ORFs. Relative abundances of transcriptionally active RTs among all transcripts varied from 0.002% (St TARA_007/DCM/180–2000 µm) to 0.15% (St TARA_007/DCM/20–180 µm) (Figure 3). Taxonomic annotation of these transcribed RTs indicated that many originated from eukaryotes (Supplementary Figure S7). However, metatranscriptomic RTs showed a distinct classification pattern to that observed in metagenomes. First, the relative abundance of LINE and BEL RTs decreased in the metatranscriptomes relative to the metagenomes, whereas the opposite was true for Gypsy (detected in 22 samples) and Copia (detected in all samples) RTs (Figure 3). Second, the relative abundances of bacterial RTs increased in the metatranscriptomic data for all fraction sizes. Group II intron RTs, which were mainly detected in pico- and piconano-fractions of metagenomes, were detected in 20 metatranscriptomes, including those from larger size fractions. Retron-type RT transcripts were detected in the metatranscriptomes from all the analyzed size fractions from Station 7. RT sequences similar to DGRs were identified in samples from Station 30. Third, the frequency of retrovirus RTs was much higher than in the metagenomes (Figure 3 and Supplementary Figure S7). These viral RTs were observed in various size fractions (22 samples). In the St TARA_023/SUR/20–180 µm sample, retroviral RTs were the most dominant type of transcribed RTs (Figure 3). Putative taxonomies were assigned for 2050 out of 4258 metatranscriptomic RT-like sequences using BLAST. Gypsy and Copia are often best matching to Tracheophyta (923 of 1092 taxonomically assigned Gypsy sequences and 388 of 463 taxonomically assigned Copia sequences), retroviruses to Euteleostomi and Retroviridae (157 and 188 out of 367, respectively), Caulimovirus to Magnoliophyta (22 of 30), Group II intron to Gammaproteobacteria (28 of 66), DGR to *Achromobacter* (10 of 10) and Retron to *Afipia* (seven of seven).

Correlation with environmental variables

The observed RT transcription pattern (Figure 4) was analyzed with DCA to identify its possible association with environmental variables (that is, 22 abiotic parameters and 107 taxon composition

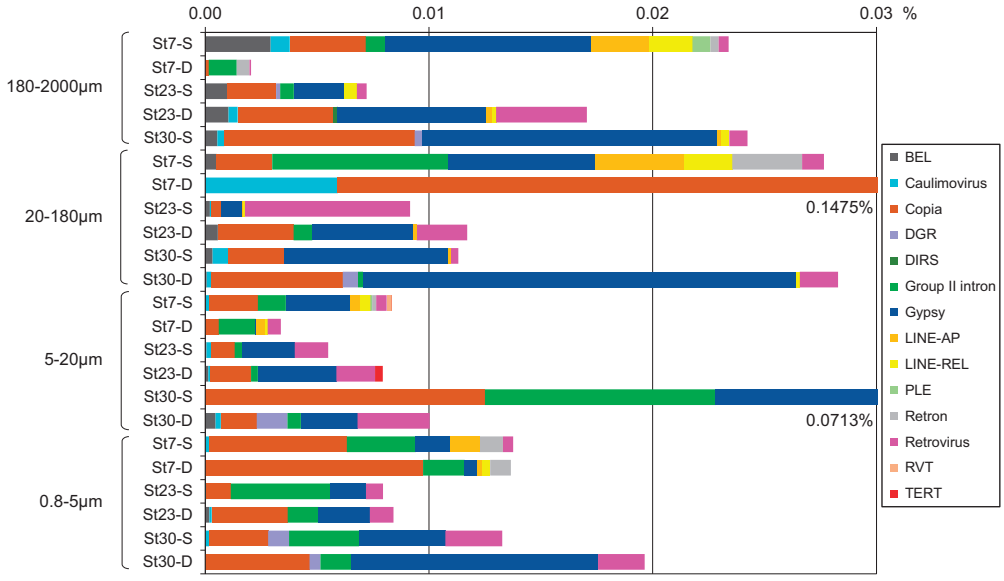

Figure 3 Relative RT gene abundance and classification of transcriptionally active RT-like sequences identified in the metatranscriptomic data.

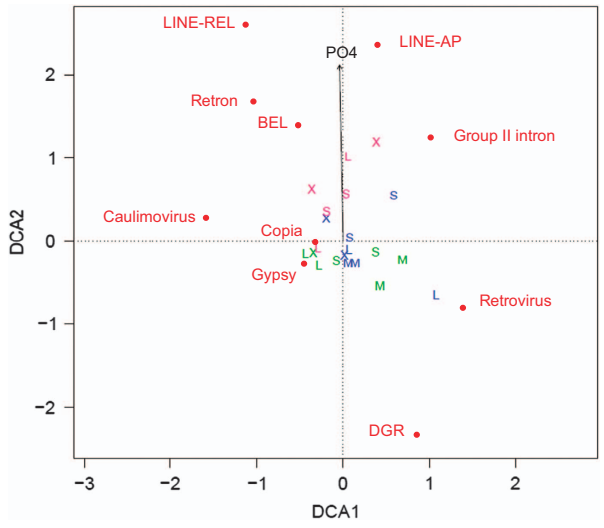

Figure 4 Detrended Correspondence Analysis (DCA) of metatranscriptomic RT gene abundance. DCA ordinations of 21 samples are shown for metatranscriptomic RT gene abundance, with significant ($P \leqslant 0.05$) environmental vectors fitted using envfit (Oksanen *et al.*, 2007). Arrows indicate the direction of the (increasing) environmental gradient, and their lengths are proportional to their correlations with the ordination. X stands for samples from size fraction 180–2000 µm, L for 20–180 µm, M for 5–20 µm and S for 0.8–5 µm. Samples from St TARA_007 is colored in pink, St TARA_023 in blue, and St TARA_030 in green.

vectors for eukaryotes). Overall, samples were clustered according to their sampling sites (rather than by size fractions) on the ordination plane. Vector fitting indicated that PO_4 concentration most significantly correlated with the RT transcription pattern ($r^2 = 0.7349$, P-value = 0.001; Figure 4 and Supplementary Table S7). Samples from St TARA_007 were placed in the region on the ordination plane where relatively high PO_4 concentrations were measured. The LINEs, BEL and retrons were abundant in these samples. When the same method was used with the metagenomic RT gene abundance, samples from the same size fraction tended to cluster together, with the 0.8–5 µm size fraction samples being outliers (Supplementary Figure S8). PO_4 ($r^2 = 0.3080$, P-value = 0.042) and NO_2+NO_3 ($r^2 = 0.3604$, P-value = 0.035) concentrations were found to be the strongest abiotic variables explaining this RT gene distribution (Supplementary Figure S8 and Supplementary Table S8). In addition, 27 taxon abundances were found to be important factors explaining the distribution, with the 0.8–5 µm size fraction samples being placed in the same direction on the ordination plane (Supplementary Table S8). The group II intron RT class strongly correlated with these biotic variables. In contrast, retron RT abundance showed a correlation with NO_2+NO_3 concentration. Overall, these analyses identified nutrients (PO_4 and NO_3) as the main environmental factors explaining the distributions of RT genes/transcripts in these Mediterranean Sea metagenomes and metatranscriptomes.

Discussion

We investigated the abundance, classification and transcriptional status of RTs using a newly generated set of marine metagenomes and metatranscriptomes. We showed a remarkably high representation of metagenomic RTs for the three largest organismal size fractions enriched in eukaryotic genomes. They were more abundant than any other known sequence domains defined in NCBI/CDD. We showed that most of these metagenomic RTs originated from retrotransposons and comprised diverse RT lineages.

The identified metagenomic RTs covered a wide spectrum of eukaryotic lineages. Our conservative taxonomic assignments suggest their presence in metazoans, stramenopiles, cryptophytes, haptophytes, Viridiplantae, Alveolata and Amoebozoa, with the most numerous assignments being to metazoans (Supplementary Figure S5). This wide phylogenetic coverage is consistent with the notion that retrotransposons are ubiquitous in eukaryotic genomes and have participated in their evolution since the radiation of the major eukaryotic lineages. Classification of the metagenomic RT sequences showed that the majority of them belonged to one of the four major known groups of retrotransposons, namely LINEs (non-LTR retroelements), Gypsy (LTR retroelements), BEL (LTR retroelements) and Copia (LTR retrotransposons). Our phylogenetic analyses revealed many previously unrecognized clades across these groups of retrotransposons (Figure 1). Furthermore, the longest RT-like ORF was identified as a cryptophyte *Dualen* element. To our knowledge, there is no previous report on the presence of *Dualen* in cryptophytes. In a previous study, Maumus *et al.*, (2009) examined the abundance of RTs in metagenomic data for size fractions under 20 μm (that is, 0.1–0.8 μm, 0.22–0.8 μm, 0.8–3.0 μm, 3.0–20.0 μm). They found a positive correlation between the frequency of RTs and the size of organisms. Our results, extending the coverage of organism sizes up to 2000 μm, revealed a similar trend (Figure 2) probably owing to an increased tolerance for transposon proliferation in large genomes of larger organisms. It should be noted that the taxonomic assignment of environmental RT sequences analyzed in the present study should be periodically revisited in the future, as more eukaryotic genomes, presently underrepresented, are added to reference databases.

RTs are known in prokaryotes but are less abundant than in eukaryotes. Well-described prokaryotic RTs are found in group II introns, retrons and DGRs (Simon and Zimmerly, 2008). We detected these three types of RTs in the metagenomic data sets from small size fractions (0.2–5 μm). Among these three types of elements, only group II introns are known to exhibit autonomous mobility (Simon and Zimmerly, 2008). Retrons are chromosomally encoded RT-encoding elements that produce multicopy single-stranded DNA/RNA molecules in the cell (Inouye *et al.*, 2011). Very little is known about their function, mobility and effect on the host cell, but retrons may be more abundant in environmental bacteria than previously recognized from the studies of model organisms given the detection of their sequences in our study as well as in a recent environmental sequencing study (Labonte *et al.*, 2015). DGRs, found in phage and bacterial genomes, can confer selective advantages to their hosts by diversifying DNA sequences through RT-mediated processes that introduce nucleotide substitutions at defined locations within a target gene (Medhekar and Miller, 2007; Miller *et al.*, 2008; Arambula *et al.*, 2013). In stark contrast to eukaryotic elements, these bacterial retroelements are not known to accumulate in high numbers within a genome (<1% of a genome) (Simon and Zimmerly, 2008). Consistently, RT-like sequences detected in the small size fraction metagenomes (0.2–5 μm) amounted only up to 0.26%. The eukaryotes whose abundance was found to correlate with the abundance of group II intron RTs might have some ecological links with bacteria harboring this type of RT (Supplementary Table S8).

The RT transcriptional landscape obtained from metatranscriptomes was markedly different from that observed from the metagenomes. First, relative RT abundances in metatranscriptomes were much lower (0.002–0.148%; 0.021% on average) than in metagenomes (0.001–13.47%; 2.23% on average) and exhibited less variability across different size fractions. This lower abundance of expressed RTs is probably due to the small proportion of transcriptionally active RT genes among those found in metagenomes. Second, the transcribed RTs were dominated by the three groups of LTR retrotransposons (Gypsy, Copia and retrovirus) and group II introns. Many of the taxonomically assigned Gypsy and Copia sequences were associated to Streptophyta, retroviruses to Vertebrata (or Retroviridae) and Group II intron to Gammaproteobacteria. Interestingly, RT transcripts from these elements were detected in all analyzed size fractions. Transcription of LINEs was previously shown to be induced by stress (Teneng *et al.*, 2007), and the detection of LINE transcripts was indeed restricted to a smaller number of samples with high PO_4 values than for LTR retrotransposons. Furthermore, we noticed that for certain bacterial RT groups, activity was confined to specific sampling stations. Retron activity was observed only in St TARA_007, and DGR only in St TARA_030, and they were often taxonomically assigned to *Afipia* and *Achromobacter*, respectively. When detected, these two groups of bacterial RTs were transcriptionally active in all the analyzed size fractions. Presence of RT transcripts from bacterial elements (group II intron, retron and DGR) in large-size fractions is likely due to the existence of bacteria physically associated to larger organisms (feeding, parasitic, endosymbiotic; (Lima-Mendez *et al.*, 2015)) or those naturally forming aggregates, for instance, in the pellet of animals (Frada *et al.*, 2014) or on the surface of microplastics. RT transcripts might be transcribed in such bacteria. It is widely accepted that many small-sized organisms can be associated with larger material or organisms (Arnosti *et al.*, 2012). Nevertheless, we cannot formally exclude the possibility of small amount of contamination of free floating bacteria. Bacteria may form artifactual aggregates during sampling and could be retained on the filters, although filter clogging was avoided in our sampling. Sequences of bacterial origin were included in our metatranscriptomic data likely due to the interference of short polyA stretches (Dreyfus and Regnier, 2002) mimicking the presence

of eukaryotic mRNA polyA-tails. Therefore, the relative abundance of bacterial RT transcripts might have been underestimated in our study.

Following a systematic analysis of the available genomes of eukaryotes, prokaryotes and viruses as well as metagenomes, Aziz *et al.* (2010) found that Class II transposases were the most abundant and ubiquitous genes in nature. However, in our list of highly abundant sequence domains (Supplementary Table S6), we found only a few such domains (for example, Tn3 transposase DDE domain (pfam01526) ranked 78th (180–2000 μm fraction), and the putative transposase OrfB (PHA02517) ranked 74th (5–20 μm fraction)). This may be due to the fact that bacterial, archaeal and viral sequences were overrepresented in the data analyzed by Aziz *et al.*, 2010. Alternatively, the Class II transposases may be underrepresented in the CDD database; Aziz *et al.* re-annotated all TEs in their study. More precise and systematic quantitative comparisons of the relative abundances of Class I and Class II TEs (enriched in different size fractions) will require other types of data than those used in these and our studies, because size-fractionated samples are not suitable for such a comparison. Therefore, we did not compare the abundance of Class II TEs with that of Class I TEs in this study. Furthermore, the high abundance of RTs and their transcription patterns observed in our environmental sequence data set from the Mediterranean Sea will have to be investigated with more global data converging a wider range of oceanic biomes.

The selfish nature of RT-encoding elements and their intricate functional/evolutionary interactions with their hosts have contributed to their success in persisting and propagating in the genomes of all domains of life. Given the abundance and transcriptional activities observed in this and previous studies, RTs in marine environments might have important roles not only for long-term evolution, but also for adaptive processes occurring within plankton populations over shorter time scales.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We are grateful to Dr Irina R Arkhipova for providing some reference RT sequences used in this study. Computational resources were provided by the Bioinformatics Center and the Supercomputer System, Institute for Chemical Research, Kyoto University. We thank the PACA-bioinfo platform for computing services. This article is contribution number 31 of the *Tara* Oceans Expedition 2009–2013. We thank the commitment of the following people and sponsors who made the *Tara* Oceans Expedition 2009–2012 possible: CNRS in particular Francoise Gaill and the Groupement de Recherche GDR3280, European Molecular Biology Laboratory (EMBL), Génoscope/CEA, the French Government 'Investissements d'Avenir' programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-11-IDEX-0001-02), the ANR projects FRANCE GÉNOMIQUE (ANR-10-INBS-09-08), POSEIDON (ANR-09-BLAN-0348), PROMETHEUS (ANR-09-GENM-031), TARA-GIRUS (ANR-09-PCS-GENM-218), PHYTBACK (ANR-2010-1709-01), SAMOSA/ANR-13-ADAP-0010, EU FP7 (MicroB3/No.287589, MaCumba/No.311975, IHMS/HEALTH-F4-2010-261376, MetaCardis/HEALTH-F4-2012-305312), ERC Advanced Grant Award to CB (Diatomite: 294823), JSPS KAKENHI (Grant Number 26430184) to HO, Agnès b, the Veolia Environment Foundation, Region Bretagne, World Courier, Illumina, Cap L'Orient, the EDF Foundation EDF Diversiterre, FRB, the Prince Albert II de Monaco Foundation, Etienne Bourgois, the *Tara* schooner and its captain and crew. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org).

Author contributions

ML performed most of the bioinformatics analyses and generated the initial version of the manuscript. PH contributed to the design of the project, to part of bioinformatics analyses and to the writing of the manuscript. KKK contributed to part of bioinformatics analyses and to the writing of the manuscript. EV contributed to part of statistical analyses. AV contributed to the analysis of Copia retrotransposons. MB contributed to the analysis of Copia retrotransposons. OJ contributed to the sequencing, assembly and annotation of the metagenomic and metatranscriptomic sequences, as well as to the writing of the manuscript. DI contributed to part of statistical analyses and characterization of sampling sites. CB contributed to the analysis of Copia retrotransposons and to the writing of the manuscript. PW contributed to the sequencing, assembly and annotation of the metagenomic and metatranscriptomic sequences, as well as to the writing of the manuscript. JMC contributed to part of bioinformatics analyses and to the writing of the manuscript. HO coordinated the project, and contributed to part of bioinformatics analyses and to the writing of the manuscript.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Antoine D, Morel A, Andre J. (1995). Algal pigment distribution and primary production in the eastern Mediterranean as derived from Coastal Zone Color Scanner observations. *J Geophys Res* **100**: 16193–16209.

Arambula D, Wong W, Medhekar BA, Guo H, Gingery M, Czornyj E *et al.* (2013). Surface display of a massively variable lipoprotein by a Legionella diversity-generating retroelement. *Proc Natl Acad Sci USA* **110**: 8212–8217.

Arkhipova IR, Pyatkov KI, Meselson M, Evgen'ev MB. (2003). Retroelements containing introns in diverse invertebrate taxa. *Nat Genet* **33**: 123–124.

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH *et al.* (2004). The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science* **306**: 79–86.

Arnosti C, Fuchs BM, Amann R, Passow U. (2012). Contrasting extracellular enzyme activities of particle-associated bacteria from distinct provinces of the North Atlantic Ocean. *Front Microbiol* **3**: 425.

Aziz RK, Breitbart M, Edwards RA. (2010). Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res* **38**: 4207–4217.

Bennetzen JL, Wang H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu Rev Plant Biol* **65**: 505–530.

Berland BR, Bonin DJ, Maestrini SY. (1980). Azote ou phosphore? Considerations sur le 'paradoxe nutrionnel' de la mer Mediterranee. *Oceanol Acta* **3**: 135–142.

Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A *et al.* (2010). The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell 22*, 2943–2955.

Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD *et al.* (2012). The genome of the polar eukaryotic microalga Coccomyxa subellipsoidea reveals traits of cold adaptation. *Genome Biol* **13**: R39.

Bork P, Bowler C, de Vargas C, Gorsky G, Karsenti E, Wincker P. (2015). Tara Oceans. Tara Oceans studies plankton at planetary scale. Introduction. *Science* **348**: 873.

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A *et al.* (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* **456**: 239–244.

Bringaud F, Muller M, Cerqueira GC, Smith M, Rochette A, El-Sayed NM *et al.* (2007). Members of a large retroposon family are determinants of post-transcriptional gene expression in Leishmania. *PLoS Pathog* **3**: 1291–1307.

Bulman S, Candy JM, Fiers M, Lister R, Conner AJ, Eady CC. (2011). Genomics of biotrophic, plant-infecting plasmodiophorids using in vitro dual cultures. *Protist* **162**: 449–461.

Carr M, Nelson M, Leadbeater BS, Baldauf SL. (2008). Three families of LTR retrotransposons are present in the genome of the choanoflagellate Monosiga brevicollis. *Protist* **159**: 579–590.

Casacuberta E, Gonzalez J. (2013). The impact of transposable elements in environmental adaptation. *Mol Ecol* **22**: 1503–1517.

Chaffron S, D'Ovidio F, Sunagawa S, Acinas SG, Coelho LP, De Monte S *et al.* (2014a), Contextual biodiversity data of selected samples from the Tara Oceans Expedition (2009–2013).

Chaffron S, Guidi L, D'Ovidio F, Speich S, Audic S, De Monte S *et al.* (2014b), Contextual environmental data of selected samples from the Tara Oceans Expedition (2009–2013).

Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J *et al.* (2014). Structural and functional partitioning of bread wheat chromosome 3B. *Science* **345**: 1249721.

Collen J, Porcel B, Carre W, Ball SG, Chaparro C, Tonon T *et al.* (2013). Genome structure and metabolic features in the red seaweed Chondrus crispus shed light on evolution of the Archaeplastida. *Proc Natl Acad Sci USA* **110**: 5247–5252.

D'Ortenzio F, Ribera d'Alcala M. (2009). On the trophic regimes of the Mediterranean Sea: a satellite analysis. *Biogeosciences* **6**: 139–148.

Darriba D, Taboada GL, Doallo R, Posada D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**: 1164–1165.

de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R *et al.* (2015). Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1261605.

Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F *et al.* (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* **36**: W465–W469.

Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, Robbens S *et al.* (2006). Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. *Proc Natl Acad Sci USA* **103**: 11647–11652.

Doolittle WF, Sapienza C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**: 601–603.

Dreyfus M, Regnier P. (2002). The poly(A) tail of mRNAs: bodyguard in eukaryotes, scavenger in bacteria. *Cell* **111**: 611–613.

Dunlap KA, Palmarini M, Varela M, Burghardt RC, Hayashi K, Farmer JL *et al.* (2006). Endogenous retroviruses regulate periimplantation placental growth and differentiation. *Proc Natl Acad Sci USA* **103**: 14390–14395.

Finn RD, Clements J, Eddy SR. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**: W29–W37.

Frada MJ, Schatz D, Farstey V, Ossolinski JE, Sabanay H, Ben-Dor S *et al.* (2014). Zooplankton may serve as transmission vectors for viruses infecting algal blooms in the ocean. *Curr Biol* **24**: 2592–2597.

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D *et al.* (2003). The genome sequence of the filamentous fungus Neurospora crassa. *Nature* **422**: 859–868.

Gifford WD, Pfaff SL, Macfarlan TS. (2013). Transposable elements as genetic regulatory substrates in early development. *Trends Cell Biol* **23**: 218–226.

Gladyshev EA, Arkhipova IR. (2011). A widespread class of reverse transcriptase-related cellular genes. *Proc Natl Acad Sci USA* **108**: 20311–20316.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.

Hingamp P, Grimsley N, Acinas SG, Clerissi C, Subirana L, Poulain J *et al.* (2013). Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J* **7**: 1678–1695.

Hua-Van A, Le Rouzic A, Boutin TS, Filee J, Capy P. (2011). The struggle for life of the genome's selfish architects. *Biol Direct* **6**: 19.

Inouye K, Tanimoto S, Kamimoto M, Shimamoto T, Shimamoto T. (2011). Two novel retron elements are replaced with retron-Vc95 in Vibrio cholerae. *Microbiol Immunol* **55**: 510–513.

Jaillon P, Chica C, Novoa EM, Hingamp P, Pelletier E, Poulain J *et al.* (submitted). Marine plankton reveal a large unknown portion of the eukaryotic gene repertoire.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**: 462–467.

Kapitonov VV, Tempel S, Jurka J. (2009). Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene* **448**: 207–213.

Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J *et al.* (2011). A holistic approach to marine eco-systems biology. *PLoS Biol* **9**: e1001177.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA *et al.* (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol* **12**: e1001889.

Khan H, Kozera C, Curtis BA, Bussey JT, Theophilou S, Bowman S *et al.* (2007). Retrotransposons and tandem repeat sequences in the nuclear genomes of cryptomonad algae. *J Mol Evol* **64**: 223–236.

Kojima KK, Fujiwara H. (2005). An extraordinary retrotransposon family encoding dual endonucleases. *Genome Res* **15**: 1106–1117.

Kojima KK, Kanehisa M. (2008). Systematic survey for novel types of prokaryotic retroelements based on gene neighborhood and protein architecture. *Mol Biol Evol* **25**: 1395–1404.

Korf I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* **5**: 59.

Krom MD, Kress N, Brenner S, Gordon L. (1991). Phosphorous limitation of primary productivity in the Eastern Mediterranean Sea. *Limnol Oceanogr* **36**: 424–432.

Labonte JM, Field EK, Lau M, Chivian D, Van Heerden E, Wommack KE *et al.* (2015). Single cell genomics indicates horizontal gene transfer and viral infections in a deep subsurface Firmicutes population. *Front Microbiol* **6**: 349.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F *et al.* (2015). Ocean plankton. Determinants of community structure in the global plankton interactome. *Science* **348**: 1262073.

Llorens C, Futami R, Covelli L, Dominguez-Escriba L, Viu JM, Tamarit D *et al.* (2011). The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res* **39**: D70–D74.

Lorenzi H, Thiagarajan M, Haas B, Wortman J, Hall N, Caler E. (2008). Genome wide survey, discovery and evolution of repetitive elements in three Entamoeba species. *BMC Genomics* **9**: 595.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J *et al.* (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**: 18.

Majorek KA, Dunin-Horkawicz S, Steczkiewicz K, Muszewska A, Nowotny M, Ginalski K *et al.* (2014). The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. *Nucleic Acids Res* **42**: 4160–4179.

Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC *et al.* (2013). CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res* **41**: D348–D352.

Maumus F, Allen AE, Mhiri C, Hu H, Jabbari K, Vardi A *et al.* (2009). Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics* **10**: 624.

Medhekar B, Miller JF. (2007). Diversity-generating retroelements. *Curr Opin Microbiol* **10**: 388–395.

Miller JL, Le Coq J, Hodes A, Barbalat R, Miller JF, Ghosh P. (2008). Selective ligand recognition by a diversity-generating retroelement variable protein. *PLoS Biol* **6**: e131.

Moreau H, Verhelst B, Couloux A, Derelle E, Rombauts S, Grimsley N *et al.* (2012). Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome Biol* **13**: R74.

Noguchi H, Park J, Takagi T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* **34**: 5623–5630.

Oksanen J, Kindt R, Legendre P, O'Hara B, Stevens MHH, Oksanen MJ *et al.* (2007). The vegan package. Community ecology package.

Orgel LE, Crick FH. (1980). Selfish DNA: the ultimate parasite. *Nature* **284**: 604–607.

Palenik B, Grimwood J, Aerts A, Rouze P, Salamov A, Putnam N *et al.* (2007). The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci USA* **104**: 7705–7710.

Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G *et al.* (2015). Open science resources for the discovery and analysis of Tara Oceans data. *Scientific data* **2**: 150023.

Piednoel M, Donnart T, Esnault C, Graca P, Higuet D, Bonnivard E. (2013). LTR-retrotransposons in R. exoculata and other crustaceans: the outstanding success of GalEa-like copia elements. *PLoS One* **8**: e57675.

Rho M, Schaack S, Gao X, Kim S, Lynch M, Tang H. (2010). LTR retroelements in the genome of Daphnia pulex. *BMC Genomics* **11**: 425.

Rice P, Longden I, Bleasby A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–277.

Riordan JD, Dupuy AJ. (2013). Domesticated transposable element gene products in human cancer. *Mob Genet Elements* **3**: e26693.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S *et al.* (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112–1115.

Schulz MH, Zerbino DR, Vingron M, Birney E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**: 1086–1092.

Simon DM, Zimmerly S. (2008). A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res* **36**: 7219–7229.

Siokou-Frangou I, Christaki U, Mazzocchi MG, Montresor M, Ribera d'Alcalá M, Vaqué D *et al.* (2010). Plankton in

the open Mediterranean Sea: a review. *Biogeosciences* **7**: 1543–1586.
Siomi MC, Sato K, Pezic D, Aravin AA. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* **12**: 246–258.
Slotkin RK, Martienssen R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Rev Genet* **8**: 272–285.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30**: 2725–2729.
Tara Oceans Consortium, Coordinators; Tara Oceans Expedition, Participants. (2014). Registry of selected samples from the Tara Oceans Expedition (2009–2013).
Teneng I, Stribinskis V, Ramos KS. (2007). Context-specific regulation of LINE-1. *Genes Cells* **12**: 1101–1110.
Toro N, Nisa-Martinez R. (2014). Comprehensive phylogenetic analysis of bacterial reverse transcriptases. *PLoS One* **9**: e114083.
Watanabe T, Nozawa T, Aikawa C, Amano A, Maruyama F, Nakagawa I. (2013). CRISPR regulation of intraspecies diversification by limiting IS transposition and intercellular recombination. *Genome Biol Evol* **5**: 1099–1114.
Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B *et al*. (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**: 973–982.
Wootton JC. (1994). Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* **18**: 269–285.
Zerbino DR, Birney E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–829.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)