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, Patrick Wincker, Jean-Michel Claverie, Hiroyuki Ogata

Supplementary Methods and Results

Extraction of DNA and RNA

Both DNA and RNA were extracted from the filters used to collect plankton corresponding to four size fractions (0.8-5 µm, 5-20 µm, 20-180 µm, 180-2000 µm). These filters were cryo-grinded in the presence of Lysis Buffer RA1 (Macherey Nagel, Germany) and β mercapto-ethanol. NucleoSpin RNA L and NucleoSpin RNA/DNA Buffer kits (Macherey Nagel) was used to purify DNA. Lysates were filtrated using a NucleoSpin Filter L. The solution was centrifuged to eliminate solid remnants. Ethanol 70% was then added to the filtrates. Filtrate/ethanol mixes were loaded onto a NucleoSpin RNA L column and washed twice with DNA wash solution. DNA was eluted with DNA elute solution and then stored in sterile microtubes at −20°C. For two samples (TARA_007/SUR and DCM/5-20 µm), DNA were amplified with Phi29 DNA polymerase based on the procedure in the Illustra GenomiPhiHYDNA Amplification kit (Epicentre Biotechnologies, USA).

For RNA purification, the remaining DNA was digested on the previous DNA NucleoSpin RNA L column with 25 µL of rDNase and 235 µL of reaction buffer for rDNase. After 15 min of incubation at room temperature, columns were washed with RA2 and RA3 buffers. RNA was eluted using RNase free water and then stored in sterile microtubes at −80°C.

DNA extraction from the filters used to collect plankton corresponding to the smallest size fraction (0.2-1.6 µm) was performed in the following way. Filters were cut into small pieces and soaked in 3 mL of lysis buffer (40 mM EDTA, 50 mM Tris-HCl, 0.75 M sucrose). Lysozyme (1 mg mL^{-1} final concentration) was added, and samples were incubated at 37°C for 45 min while slightly shaken. Sodium dodecyl sulfate (1% final concentration) and proteinase K (0.2 mg mL^{-1} final concentration)

were then added. These samples were incubated at 55°C for 60 min while slightly shaken. Two rounds of phenol-chloroform extraction were performed for the lysate to extract genomic DNA.

cDNA synthesis

cDNA synthesis was carried out with the use of 100 to 200 ng of purified mRNA. The MicroPoly(A)Purist kit (Ambion, USA) was used to isolate mRNA and the CloneMiner II kit (Invitrogen) was used to constructed cDNA libraries. First-strand cDNA synthesis was performed with the RNA oligonucleotide biotin-attB2-oligo dT and SuperScript III reverse transcriptase. After second strand synthesis and blunt-end repair, the phosphorylated double-stranded attB1 adapter was ligated to the 5'-end of the cDNA. The BDAdvantage 2 PCR kit and the Att-u (CCGCGCGCACAACTTTGTAC) oligonucleotide were used to perform 8 to 12 cycles of PCR amplification.

Sequencing

Metagenomic and metatranscriptomic libraries were prepared using Illumina's protocol. For DNA samples, 30 to 50 ng of DNA was sonicated to 100 to 800 bp using the E210 Covaris instrument (Covaris, Inc., USA); for cDNA, 30 ng was sonicated to 150 to 600 bp with the same instrument. For both DNA and cDNA, fragments were end-repaired and 3'-adenylated, and Illumina adapters were added. Fragments were PCR-amplified using Illumina adapter specific primers and purified. Finally, libraries were quantified by qPCR (MxPro, Agilent Technologies), and libraries profiles evaluated with an Agilent 2100 Bioanalyzer (MxPro, Agilent Technologies). Each library was sequenced using 101 bp long read chemistry in a paired-end flow cell on Illumina sequencer (Illumina, USA) to obtain overlapping reads for generating 180 bp long merged reads.

Copy number index

In this study, the relative gene abundances in the metagenomic and metatranscriptomic data sets were estimated by their "average contig coverages" defined by the cumulative sizes of mapped reads on the contigs divided by their sizes.

To assess the quantitative nature of the average contig coverage, we independently computed the copy number index (CNI), which we defined as follows: the density of reads mapped on RT-related CDD position-specific score matrices (PSSMs) divided by the median of the density of reads mapped on CDD profiles for single copy genes. More precisely, we first identified ORFs in merged paired-end reads and prepared a set of amino acid sequences (≥40 aa). Low complexity regions within these ORFs were filtered out using SEG. CDD profiles related to RTs (34 PSSMs: pfam00078, cd01650, cd01647, cd01651, cd09274, cd03715, cd01644, cd01645, cd09076, cd03714, cd00303, cd01646, cd09275, pfam07727, cd05484, cd09077, cd09272, cd00304, cd05481, cd10442, cd01648, cd03487, cd09276,

pfam00075, TIGR00195, COG3344, cd06094, cd09273, pfam13966, KOG1005, cd00719, pfam13456, cd01438, cd06095) and those for single copy marker genes (35 PSSMs, (1)) were searched against the read-derived ORF set using PSI-BLAST (E-value < 0.001). The number of reads with significant sequence similarity to each RT-related PSSM was then divided by the average length of sequences used in the PSSM (for size normalization). The size normalized abundance was further divided by the median of the size normalized abundances of single copy marker genes to obtain a CNI value for the RT-related PSSM. This procedure defined the CNI values for RT-related PSSMs based on read data. The CNI values based on the assembly data were calculated in a similar way starting from the predicted ORFs derived from contigs with a normalization by the read density of single copy marker genes, except that we estimated the density of reads associated to the predicted amino acid sequences using the average contig coverages. The CNI values were computed for the read and assembly data for the metagenome from St23/DCM/180-2000 μm, from which the largest number of RT-related ORFs were identified.

Assessment of relative abundance measure

Relative abundances of RTs in metagenomes and metatranscriptomes were measured using the average contig coverage, which were computed for genes (or gene fragments) located in contigs longer than 500 bp. Since RT sequences could exist as multiple, nearly identical copies in the genomes collected in the samples, RT genes might have benefited from better chance of assembly (thus a higher representation in the resulting contig sets) than other genes with fewer numbers of copies. Therefore, these abundance estimates might have led to an overestimation of RTs relative to other types of genes with similar abundances but with higher sequence divergence. To examine this possibility, we analyzed the unassembled high quality reads from the St23/DCM/180-2000 μm sample (5.6 million reads, average length 173 bp), and compared the assembly versus read based copy number index (CNI) values for 34 RT-related CDD domains. If the abundance of RT-like sequences were overestimated in the previous analysis based on contigs, we would have expected larger CNI values for the assembly data than those for the read data. The result revealed no such tendency (**Figure**). Although there were several outliers, the correlation of CNI values between assembly and read datasets was significant (Pearson's r=0.68, p=1.1x10-5). The CNI values were larger for the assembly data than for the read data for 14 CDD domains, and the remaining RT domains (20 CDDs) showed a reverse tendency (two-sided binomial test, p=0.39; non-significant). The slope of a regression line crossing the origin was 1.98 (i.e., y=1.98x), indicating higher CNI values in the read-based analysis. Therefore, we concluded that the abundance estimate based on contigs was a reasonable (slightly conservative) proxy for the relative abundances of RT-like sequences in the samples.

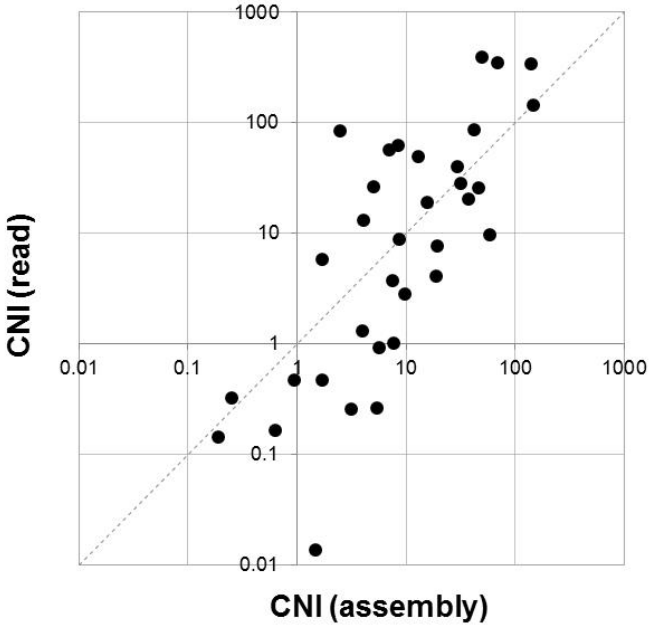

Figure. Comparison of copy number index (CNI) for RT-related CDDs. The analyzed data are from the St23/DCM/180-2000µm sample. Relatively good 1:1 ratio (especially those CDD showing high abundances) between the estimate from the assembly and those from the raw reads indicate that the abundance estimate from the assembly can be considered quantitative.

Reference

Raes, J., Korbel, J.O., Lercher, M.J., von Mering, C. and Bork, P. (2007) Prediction of effective genome size in metagenomic samples. *Genome Biol*, **8**, R10.