Comparing PCR-generated artifacts of different polymerases for improved accuracy of DNA metabarcoding

Satoshi Nagai¹, Sirje Sildevaer¹,², Noriko Nishi³, Satoshi Tazawa³, Leila Basti⁴, Takanori Kobayashi¹, Yoshizumi Ishino⁵

¹ Research Center for Bioinformatics and Biosciences, National Research Institute of Fisheries Science, 2-12-4 Fukuura, Kanazawa-ku, Yokohama, Kanagawa 236-8648, Japan
² Department of Marine Systems, Tallinn University of Technology, Akadeemia tee 15A, 12618 Tallinn, Estonia
³ AXIOHELIX Co. Ltd, 1-12-17 Kandaizumicho, Chiyoda-ku, Tokyo 101-0024, Japan
⁴ Department of Ocean Sciences, Faculty of Marine Resources and Environment, Tokyo University of Marine Science and Technology, 4-5-7 Konan, Minato-ku, Tokyo 108-8477, Japan
⁵ Department of Bioscience and Biotechnology, Graduate School of Bioresource and Bioenvironmental Sciences, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan

Corresponding author: Satoshi Nagai (snagai@affrc.go.jp)

Abstract

Accuracy of PCR amplification is vital for obtaining reliable amplicon-sequencing results by metabarcoding. Here, we performed a comparative analysis of error profiles in the PCR products by 14 different PCR kits using a mock eukaryotic community DNA sample mimicking metabarcoding analysis. To prepare a mock eukaryotic community from the marine environment, equal amounts of plasmid DNA from 40 microalgal species were mixed and used for amplicon-sequencing by a high-throughput sequencing approach. To compare the differences in PCR kits used for this experiment, we focused on the following seven parameters: 1) Quality, 2) Chimera, 3) Blast top hit accuracy, 4) Deletion, 5) Insertion, 6) Base substitution and 7) Amplification bias amongst species. The results showed statistically significant differences (p < 0.05) for all of the seven parameters depending on the PCR kits used. These differences may result from the different DNA polymerases included in each kit, although the result can also be influenced by PCR reaction conditions. Simultaneous analysis of several parameters suggested that kits containing KOD plus Neo (TOYOBO) and HotStart Taq DNA polymerase (BiHONEER, CA, US) at the annealing temperature of 65 °C displayed better results in terms of parameters associated with chimeras, top hit similarity and deletions.

Key Words

chimeric sequences, DNA polymerase, high throughput sequencing, metabarcoding, PCR-generated artifacts, 18S rRNA gene

Introduction

DNA polymerases are widely used for DNA manipulation in vitro, including DNA cloning, sequencing, labelling, mutagenesis and other purposes. The fundamental ability of DNA polymerases to synthesise a deoxyribonucleotide chain is conserved. However, the more specific properties, including thermostability, fidelity (proofreading activity), processivity (conservation of sequential nucleotide polymerisation) and specificity (proportion of non-specific amplification) vary depending on the polymerase. For example, thermostability can range from < 4 to 1380 min at 95 °C amongst DNA polymerases (van Pelt-Verkuil et al. 2008). Some polymerases also have a 3′-5′ proofreading activity that corrects occasionally-occurring nucleotide misincorporations during the extension processes. However, some polymerases, for example, Taq DNA polymerase, lack proofreading ability, leaving the errors uncorrected (Taberlet et al. 2018). For this reason, the usage of high-fidelity DNA polymerase has been suggested for reducing erroneous
sequences obtained by DNA metabarcoding (Oliver et al. 2015; Sze and Schloss 2019). The fidelity of polymerases with proofreading ability may be up to 300 times higher compared to Taq polymerase (van Pelt-Verkuil et al. 2008; Potapov and Ong 2017; Thermofisher Scientific 2021).

Processivity indicates the number of nucleotides added to the DNA sequence during a single binding event (Wang et al. 2004). Polymerases with higher processivity support amplification of long templates and shorter extension time and lower amount of polymerase are needed for successful amplification (Wang et al. 2004). For example, KOD DNA polymerase exhibits excellent processivity, showing a five-fold higher extension rate (100–130 nucleotides/second) and 10–15-fold higher processivity (> 300 bases) than Pfu DNA polymerase (Takagi et al. 1997). Specificity of the DNA polymerases reflects the proportion of non-specific amplification, for example, extension of misprimed targets and primer-dimers that can have a notable impact on the yield and sensitivity of target amplification (Chou et al. 1992).

For these reasons, protein engineering techniques to create mutant or artificial DNA polymerases have been successfully applied for developing more powerful DNA polymerases, suitable for specific purposes amongst the different kinds of DNA manipulations (Ishino and Ishino 2014). In addition, manufacturers also provide optimised PCR kits containing polymerase and other reagents for efficient amplification of various templates.

The metabarcoding approach using universal primers in Illumina sequencers has become a popular method for environmental DNA analysis in aquatic ecosystems, owing to the development of high-throughput sequencing (HTS) technologies (Amaral-Zettler et al. 2009; Medinger et al. 2010; Edgcomb et al. 2011; Taberlet et al. 2012; Egge et al. 2013, 2015; Majaneva et al. 2015; Sawaya et al. 2019). The significant advantage of applying HTS-based technology for monitoring biodiversity offers the great potential for more precise species identification, based on genetic information, especially for species that are indistinguishable by traditional morphology-based microscopic observation (Rhodes 1998; John et al. 2005). This technology delivers high-throughput performance and allows for the detection of several hundreds of operational taxonomic units (OTUs), including dominant species and/or hidden flora from aquatic ecosystems (Cheung et al. 2010; Nolte et al. 2010; Monchy et al. 2012; Tanabe et al. 2015; Nagai et al. 2016a, b, 2019; Dzhembekova et al. 2017, 2018; Hirai et al. 2017a, b; Sildever et al. 2019).

The resulting HTS data is influenced by various factors from DNA extraction to bioinformatics data analysis (Oliver et al. 2015; Deiner et al. 2018; Nichols et al. 2018; Jeunen et al. 2019; Santoferrara 2019; van der Loos and Nijland 2020; Zaiko et al. 2021). Amongst those factors, PCR amplification can influence the diversity detected and relative sequence abundances obtained by the HTS data (Haas et al. 2011; Brandarriz-Fontes et al. 2015; Kelly et al. 2019). The influence of polymerase choice on the HTS data has been previously investigated in terms of characterisation of PCR-related errors (Quail et al. 2012; Gohl et al. 2016; Oh et al., n.d.), chimera formation (Lahr and Katz 2009), species occurrence and relative sequence abundances (Haas et al. 2011; Brandarriz-Fontes et al. 2015; Oliver et al. 2015; Nichols et al. 2018; Kawato et al. 2021) and community composition and quality of HTS data (Sze and Schloss 2019). PCR-generated artifacts have also been found to increase as species diversity increases (Qiu et al. 2001); thus, amplicon-sequence data of 16S rRNA or 18S rRNA from environmental DNAs may contain several artifacts. Furthermore, in a cross-laboratory experiment, the choice of the polymerase had a consistently significant effect on the metabarcoding data variability explained by the differential performance of the polymerases used (Zaiko et al. 2021).

Therefore, we examined the effect of the selected PCR kits containing polymerase on the accuracy of amplicon-sequence reads of the 18S rRNA gene using the HTS-based technology. A mock sample of a eukaryote community was prepared by equally pooling plasmid DNAs from 40 microalgal species. The frequency of artifacts was compared amongst 14 PCR kits containing polymerase in the following seven parameters, i.e. frequencies of: 1) Quality; 2) Chimera; 3) Blast top hit accuracy; 4) Deletion; 5) Insertion; 6) Base substitution; and 7) Amplification bias amongst species. The results displayed statistically significant differences in the amplicon sequences from different PCR kits. Various studies can utilise metabarcoding analysis and the results generated through the experiments reported here will contribute to the planning of amplicon-based HTS studies and evaluation of the obtained data in terms of PCR kit (polymerase) associated bias.

Materials and methods

Abbreviations

HTS  high-throughput sequencing;
OTUs  operational taxonomic units.

DNA sample preparation of a mock community

Clonal strains of 40 microalgal species were isolated from plankton blooms in several localities from Japan (Suppl. material 1: Table S1). All strains were maintained in glass test tubes in 6 ml according to Nagai et al. (2008). The clonal strains were individually cultivated and DNAs were extracted from the harvested cells by the procedure reported previously by Nagai et al. (2008). Universal primer pair (TAREuk454FWD1, F: CCAGCASCYGCG-GTAATTCC; TAREuk454REV3, R: ACTTTCGTTCTT-GATYRA; Stoeck et al. 2010) was used to amplify the V4 hypervariable regions of the 18S rRNA gene. PCR was performed on the thermal cycler in a reaction mixture (25 μl) containing 1.0 μl of template DNA, 0.2 mM of each dNTP, 1× PCR buffer, 1.5 mM Mg²⁺, 1.0 U of KOD -Plus- ver.2 (TOYOBO) that has intensive 3'→5' exonuclease activity and 1.0 μM of each primer. The PCR
cycling conditions were as follows: initial denaturation at 94 °C for 3 min; 30 cycles at 94 °C for 15 s, at 56 °C for 30 s and at 68 °C for 40 s. Results of PCR amplification were checked by agarose gel electrophoresis. Cloning of the amplified DNA fragments followed by sequencing was carried out as described by Nagai et al. (2008). All the 40 sequences obtained in this study are available from GenBank (accession numbers in Suppl. material 1: Table S1). The plasmid DNAs containing a DNA fragment amplified from the target species were purified from each E. coli transformant cell, cultivated in 2 ml of the LB medium at 37 °C for 18 hours, using a FastGene Plasmid Mini Kit (NIPON Genetics, Tokyo, Japan) according to the manufacturer’s instruction. The purified plasmid DNAs were quantified using a Qubit 2.0 Fluorometer (Life technology, Carlsbad, CA, USA) and the concentrations were 4.2–51.0 ng µl\(^{-1}\) (22.2 ± 11.0, Mean ± SD, n = 40). The DNA samples were pooled with an equal amount (150 ng) to the final concentration of 16.0 ng µl\(^{-1}\) and stored at -30 °C until use.

**Paired-end library preparation and MiSeq sequencing**

To carry out metabarcoding analysis using the MiSeq 250PE platform (Illumina, USA), the same universal primer pair targeting 18S rRNA gene V4 hypervariable regions (around 415 bp in length; Stoeck et al. 2010) was used for amplification due to the unavoidable restriction of sequence length. The workflow followed the “16S metagenomic sequencing library preparation: preparing 16S ribosomal gene amplicons for the Illumina MiSeq system” distributed by Illumina (part no. 15044223 Rev. B). A two-step PCR approach was employed to construct the paired-end libraries. The first-round PCR amplified the target region using primers 5′-ACACTCTTCTCCCTACACGACGCTCTTCCGATCT + 18S rRNA gene (TAReu454FWD1) and 5′-GTGACTGGAGTTCAGACGT - GTG (reverse). The eight base segments represent dual-index sequences used to recognise each sample; the 5′ end-sequences are adapters that allow the final product to bind or hybridise to short oligonucleotides on the surface of the Illumina flow cell and the 3′ end-sequences are priming sites for the MiSeq sequencing. The purified first-round PCR product was diluted 5 times with TE buffer and used as a template for the second-round PCR. The second-round PCR was carried out with the same condition as the first-round PCR, except the reaction volume of 50 µl containing 2.0 µl of the diluted PCR product. Eight cycles of the PCR with the annealing temperature of 59 °C were set for all PCR reactions tested in this study (Suppl. material 1: Table S2). PCR amplification was verified by agarose gel electrophoresis and the PCR products were purified using an Agencourt AMPure XP (BECKMAN COULTER). The amplified PCR products were quantified and the indexed second PCR products were pooled in equal concentrations and stored at -30 °C until used for sequencing.

When performing high throughput sequencing (HTS), a spike-in of the control library (Illumina standard product: PhiX V3 Control Library) was mixed with the pooled DNA library to improve the data quality of low diversity samples, such as single PCR amplicons. DNA concentrations of the pooled library and the PhiX DNA were adjusted to 4 nM using the buffer EB (10 mM Tris-HCl, pH 8.5) mixed at a ratio of 7:3.5 µl. The 4 nM library was denatured with 5 µl of fresh 0.1 N NaOH. Using the HT1 buffer (provided by the Illumina MiSeq v. 2 Reagent kit for 2× 250 bp PE), the denatured library (10 µl, 2 nM) was diluted to a final concentration of 12 pM for sequencing on the MiSeq platform.

**HTS data treatment processes and operational taxonomic unit picking**

Nucleotide sequences were demultiplexed depending on the 5′-multiplex identifier (MID) tag and primer sequences using the default format in MiSeq. The sequences containing palindrome clips longer than 30 bp and homopolymer longer than 9 bp were trimmed from the sequences at both ends. The 3′ tails with an average quality score of less than 30 at the end of the last 25 bp window were also trimmed from each sequence. The 5′ and 3′ tails with an average quality score of less than 20 at the end of the last window were also trimmed from each sequence. Sequences longer than 250 bp were truncated to 250 bp by trimming the 3′ tails. The trimmed sequences shorter than 200 bp were filtered out. The demultiplexing and trimming were purified using an Agencourt AMPure XP (BECKMAN COULTER, Life Sciences, Brea, California, USA) according to the manufacturer’s protocol.
Table 1. Overview of PCR kits compared in this study based on manufacturer’s information, numbering follows Figure 1.

<table>
<thead>
<tr>
<th>Nr.</th>
<th>PCR kit containing polymerase</th>
<th>Manufacturer</th>
<th>Proofreading ability</th>
<th>Proofreading compared to Taq</th>
<th>Hot start</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>KOD-Plus-Ver.2</td>
<td>TOYOBO</td>
<td>Y</td>
<td>80</td>
<td>Y</td>
</tr>
<tr>
<td>2–4.</td>
<td>KOD-Plus-Neo</td>
<td>TOYOBO</td>
<td>Y</td>
<td>80</td>
<td>Y</td>
</tr>
<tr>
<td>5.</td>
<td>KOD FX Neo</td>
<td>TOYOBO</td>
<td>Y</td>
<td>80</td>
<td>Y</td>
</tr>
<tr>
<td>6.</td>
<td>Go Taq Green Master Mix</td>
<td>Promega</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>Type-it Microsatellite PCR Kit</td>
<td>QIAGEN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>Q5 High-fidelity DNA Polymerase</td>
<td>NEW ENGLAND BioLabs</td>
<td>Y</td>
<td>280</td>
<td>Y</td>
</tr>
<tr>
<td>9.</td>
<td>One Taq Hot Start DNA Polymerase</td>
<td>NEW ENGLAND BioLabs</td>
<td>Y</td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>10.</td>
<td>KAPA2GRobust HotStart Ready Mix with dye (2X)</td>
<td>KAPA Biosystems</td>
<td>N</td>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>11.</td>
<td>KAPA HiFi HotStart Ready Mix</td>
<td>KAPA Biosystems</td>
<td>Y</td>
<td>100</td>
<td>Y</td>
</tr>
<tr>
<td>12.</td>
<td>Premix Ex Taq Hot Start Version</td>
<td>TaKaRa</td>
<td>Y</td>
<td>4.5</td>
<td>Y</td>
</tr>
<tr>
<td>13.</td>
<td>Top DNA Polymerase</td>
<td>BIONEER</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td>Pfu DNA Polymerase</td>
<td>BIONEER</td>
<td>Y</td>
<td>30</td>
<td>Y</td>
</tr>
<tr>
<td>15–17.</td>
<td>HotStart Taq DNA Polymerase</td>
<td>BIONEER</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18.</td>
<td>Platinum SuperFi PCR Master Mix</td>
<td>Invitrogen</td>
<td>Y</td>
<td>300</td>
<td>Y</td>
</tr>
</tbody>
</table>

* combination of Taq (non-proofreading) and Vent (proofreading) polymerases, † Vent polymerase, ‡ combination of Taq and exonuclease, 4 Potapov and Ong (2017).

Sequence comparisons obtained by PCR with various kits containing different DNA polymerases

The effect of fourteen PCR kits containing polymerase on the accuracy for amplification of 18S-rRNA gene in a mock sample of 40 microalgal species was examined in terms of performance in the following seven parameters: 1) Quality; 2) Chimera; 3) Blast top-hit accuracy; 4) Deletion; 5) Insertion; 6) Base substitution; and 7) Amplification bias amongst species. The calculation of each of the parameters is described in Table 2. Bioinformatics analysis from data processing by Trimomatic to sequence comparison was performed for each PCR condition and the DNA polymerase (n = 18). In Quality, Trimomatic file outputs were used for the count of reads numbers. Read numbers in each sample were obtained by a simple Linux command to count the line number (LN) (e.g. wc -l *_fastq) and it was divided by four because one sequence information is mentioned with four lines. In Chimera, read numbers before the Chimera check is the same as the read numbers passed in the quality check in Quality. Read numbers after the Chimera check counted by count.seqs was obtained from the Mothur file outputs. Three Perl scripts were made to analyse Blast top hit frequency, Deletion, Insertion, Base substitution, and Amplification bias. Blast XML files were read by blastxml_parser.pl (provided as Suppl. material 2) using Bio::SearchIO module in BioPerl. Results of the BLAST search were processed for each query (OTU) and only the top-hit records were extracted.

The following information: query ID, query length, bit score, top-hit name, top-hit identity, alignment length, query alignment length, query alignment sequence, reference alignment sequence and homology information between the query and reference sequences in the alignment (homology_string) were contained in a text file output. Numbers of top-hit sequences to the reference sequences of 40 species for each query ID were counted in the triplicate samples by count.seqs command in Mothur and merged with the information obtained from Blast XML by merge_tophit_count.pl (provided as Suppl. material 2).
Table 2. Overview of the seven parameters used for evaluation of PCR kits based on frequencies, PE: paired-end, SD: standard deviation.

<table>
<thead>
<tr>
<th>Order</th>
<th>Parameter</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Quality</td>
<td>Read numbers, passed quality check (the number of reads after merge assembly of PE reads following quality trimming but before chimera check) / raw read numbers in each sample</td>
</tr>
<tr>
<td>B</td>
<td>Chimera</td>
<td>Read numbers after Chimera check / read numbers before Chimera check in each sample.</td>
</tr>
<tr>
<td>C</td>
<td>Blast top hit accuracy</td>
<td>Numbers of unique hit-sequence to the reference sequences of 40 species / input sequence-numbers in Blast search (+ read numbers after Chimera check) in each sample.</td>
</tr>
<tr>
<td>D</td>
<td>Deletion</td>
<td>Deletion = 1- ((the number of deleted base in OTU1 × the read count of OTU1 + the number of deleted base in OTU2 + the number of deleted base in OTU40 + • • • + number of deleted base in OTU40) / (the alignment length of OTU1 × the read count of OTU1 + the alignment length of OTU2 × the read count of OTU2 + • • • + the alignment length of OTU40 × the read count of OTU40)) × 100</td>
</tr>
<tr>
<td>E</td>
<td>Insertion</td>
<td>Insertion = 1- ((the number of inserted base in OTU1 × the read count of OTU1 + the number of inserted base in OTU2 + the number of inserted base in OTU40 + • • • + number of inserted base in OTU40) / (the alignment length of OTU1 × the read count of OTU1 + the alignment length of OTU2 × the read count of OTU2 + • • • + the alignment length of OTU40 × the read count of OTU40)) × 100</td>
</tr>
<tr>
<td>F</td>
<td>Base substitution</td>
<td>Base substitution = 1- ((the number of base substitution in OTU1 × the read count of OTU1 + the number of base substitution in OTU2 + the number of base substitution in OTU40 + • • • + number of base substitution in OTU40) / (the alignment length of OTU1 × the read count of OTU1 + the alignment length of OTU2 × the read count of OTU2 + • • • + the alignment length of OTU40 × the read count of OTU40)) × 100</td>
</tr>
<tr>
<td>G</td>
<td>Amplification bias among species</td>
<td>1-normalized SD of numbers of unique hit-sequences to the reference sequences of 40 species for 40 species</td>
</tr>
</tbody>
</table>

nutation was saved as an output in a text file (result_merge.txt). Subsequently, numbers of unique hit-sequences to the reference sequences of 40 species were counted, based on triplicate samples by same_tophit_count_merge.pl (provided as Suppl. material 3). The information was saved as an output in a text file. Removal of sequences containing errors was imperfect after the successive processes of HTS data treatment. Sequences containing different types of errors derived from original ones remained in the following analytical steps. Therefore, these sequences are detected as unique OTUs with the same BLAST top-hit name, but different similarities. BLAST top-hit accuracy and Amplification bias amongst species were calculated using this file. Deletion, Insertion and Base substitution were calculated, based on the following command:

```
awk -F"\t" '/&NR!=1{ OFS="\t";}x[1]=gsub("-", "\", $8);x[2]=gsub("-", "\", $9) ;x[3]= gsub(/\s/, "\", $10);print $1,x[1],x[2],x[3]; }' result_merge.txt
```

Namely, to detect Deletion, Insertion and Base substitution, the number of the hyphen ‘-’ in the query alignment sequence, the number of the hyphen ‘-’ in the reference alignment sequence and the number of space between the query and reference sequences in the alignment were counted respectively and the information was saved as an output in a text file. For Amplification bias amongst species, the normalised standard deviation (SD) of numbers of unique hit sequences in 40 species were calculated for each PCR condition in Excel.

To evaluate whether there are statistically significant differences amongst the results from different PCR kits, based on the seven parameters, one-way ANOVA or Kruskal-Wallis rank-sum tests were conducted in R (R Core Team 2021). The assumptions for the tests: normal distribution (ANOVA) and equal variation within groups (ANOVA, Kruskal-Wallis) were also confirmed in R by applying Shapiro-Wilk and Levene’s test (package “car”, Fox et al. 2020) for normally distributed data or Flinger-Killeen test in the case of non-normal distribution. Tukey Honest Significant Differences method with family-wise confidence level (95%) or Dunn test with the P-value adjusted by the Benjamini-Hochberg method (package “FSA”, Ogle et al. 2021) were used as post hoc tests for statistically significant multiple comparisons to obtain more detailed information on the differences amongst the PCR kits tested in this study.

To display performances of PCR kits containing DNA polymerases amongst the investigated parameters intuitively, radar charts were employed. Radar charts represent the integrated performance evaluation of seven parameters in each PCR kit containing polymerase (14 kits) and three different annealing temperature conditions for KOD-Plus-Neo and HotStart Taq DNA Polymerase. The values on the radar charts range from 0–1; if the value is closer to 1, the PCR kit has a better performance in terms of the particular parameter.

Results

All of the seven parameters used for the comparison of the sequence data from different PCR kits displayed statistically significant differences (p < 0.05; Table 3). For example, the differences in the results of quality before/after quality trimming amongst different PCR kits and the reaction conditions ranged from 0.29 to 0.41 with higher frequencies indicating higher quality (Fig. 1A). Although the quality amongst different PCR kits varied in a narrow range, a statistically significant difference (p < 0.05; Table 3) was detected, related to Premix Ex Taq Hot Start Version (nr. 12; Fig. 1A; Suppl. material 1: Table S3A). For the parameter associated with chimeras, a notable and statistically significant variability amongst the different PCR kits was observed (Fig. 1B; Suppl. material 1: Table S3B). Significantly better performance (lower frequencies of chimeras) was demonstrated by KOD.
Table 3. Comparison between different polymerases based on the parameters evaluated.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Saphiro-Wilk (W/p value)</th>
<th>Levene's test (F/p value)</th>
<th>Flinger/Killeen (X2/p)</th>
<th>ANOVA (F/p value)</th>
<th>Kruskal-Wallis (X2/p value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality</td>
<td>0.98/0.71</td>
<td>1.64/0.99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chimera</td>
<td>0.85/6.645e-06</td>
<td>9.31/0.93</td>
<td>13.23/0.72</td>
<td>52.52/1.702e-05</td>
<td></td>
</tr>
<tr>
<td>Blast top hit accuracy</td>
<td>0.44/8.418e-13</td>
<td>12.43/0.77</td>
<td>52.18/1.925e-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deletion</td>
<td>0.92/1.323e-06</td>
<td>8.10/0.96</td>
<td>52.39/1.782e-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insertion</td>
<td>0.91/0.001</td>
<td>12.40/0.77</td>
<td>50.68/3.313e-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base substitution</td>
<td>0.93/0.003</td>
<td>5.35/0.99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amplification bias</td>
<td>0.57/2.848e-11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Plus NEO and Hot Start Taq DNA Polymerase with the annealing temperature of 65 °C (nr. 4, 17; Fig. 1B; Suppl. material 1: Table S3B). At the same time, significantly lower performance (higher frequencies of chimeras) was present, when Q5 High-fidelity DNA Polymerase and Platinum SuperFi PCR Master Mix (nr. 8, 18; Fig. 1B; Suppl. material 1: Table S3B) were employed.

The performance of reads detected against the unique reference sequences by BLAST search, parameter 3) Blast top-hit accuracy, displayed similarly high performance amongst the majority of the PCR kits, with significantly lower values obtained by KAPA2G Robust HotStart Ready Mix with dye (2X), KAPA HiFi HotStart Ready Mix, Premix Ex Taq Hot Start Version and Top DNA Polymerase (nr. 10-13; Figs 1C, 2; Suppl. material 1: Table S3C).

In terms of deletions and insertions, the variability amongst the different PCR kits was low (Fig. 1D, E). In contrast, more deletions were detected in the sequences amplified by Q5 High-fidelity DNA Polymerase and Platinum SuperFi PCR Master Mix (nr. 8, 18). In the case of insertions, significantly higher bias was detected from the DNA amplified by Go Taq Green Master Mix, Premix Ex Taq Hot Start Version and Platinum SuperFi PCR Master Mix (nr. 6, 12, 18; Fig. 1E, Suppl. material 1: Table S3E). The frequency of base substitutions also varied significantly amongst the PCR kits with the lowest performance by KOD FX Neo and KAPA2G Robust HotStart Ready Mix with dye (2X) (nr. 5, 10; Fig. 1F; Suppl. material 1: Table S3F). Higher performance (the lower numbers of base substitution) was demonstrated by KOD Plus Neo at 65 °C, Q5 High-fidelity DNA Polymerase, KAPA HiFi HotStart Ready Mix and HotStart Taq DNA Polymerase with the annealing temperature of 65 °C (nr. 4, 8, 11, 17; Fig. 1F; Suppl. material 1: Table S3F). The PCR amplification bias amongst the sequences varied significantly in a range from 0.52 to 0.74, with the significantly lowest performance obtained by Q5 High-fidelity DNA Polymerase (nr. 8; Fig. 1G; Suppl. material 1: Table S3G).

Amongst all the parameters evaluated, Premix Ex Taq Hot Start Version displayed significant differences with other PCR kits in four parameters (nr. 12; Fig. 1A–G; Suppl. material 1: Table S3A–G). For other polymerases, the frequency patterns were more variable amongst different parameters (Fig. 1, Suppl. material 1: Table S3). In terms of annealing temperatures, a significant difference for KOD Plus Neo was detected in the frequency of chimeras, top hit similarity, and deletions between 56 °C and 65 °C (nr. 2, 4; Fig. 1B, C; Suppl. material 1: Table S3).
S3B, C). For HotStart Taq DNA Polymerase, significant differences between the annealing temperatures were detected for the frequency of deletions at 56 °C and 65 °C (nr. 15, 17; Fig. 1D, Suppl. material 1: Table S3D) as well as for the amplification bias at 61 °C and 65 °C (nr. 16, 17, Fig. 1G; Suppl. material 1: Table S3G).

When considering different parameters together, the majority of PCR kits displayed high performance in two parameters, which were mainly associated with the deletions and top-hit similarities (Fig. 2). Thirteen PCR kits showed the highest results in terms of one parameter that was mainly associated with the top-hit similarity. One PCR kit (Premix Ex Taq Hot Start Version) did not show high performance in any of the parameters, whereas two kits at 65 °C (KOD Plus Neo at 65 °C and Hot Start Taq DNA polymerase at 65 °C) had the best results for the highest number of parameters, mainly associated with chimeras, top-hit similarity and deletions (Fig. 2).
Discussion

We detected significant differences amongst PCR kits for each of the compared parameters when using the mock community sample of marine microeukaryotes. The highest number of significant differences amongst kits was detected in association with BLAST top-hit accuracy, chimeras, insertions and base substitutions, whereas the lowest number of significant differences was detected in terms of the quality of sequences. The low number of significant differences in terms of quality may potentially be explained by similar sequencing depth due to equal concentrations of the pooled samples (Elbrecht and Steinke 2018). Based on all the investigated parameters, the majority of PCR kits performed well in terms of top-hit similarity and deletions.

When analysing the different parameters in more detail, the first parameter displaying a high number of significant differences amongst kits was associated with chimera formation. It varied notably amongst the PCR kits and was the highest in the samples amplified by Q5 High-fidelity DNA Polymerase and Platinum SuperFi PCR Master Mix (Fig. 1B). In a previous study, Q5 High-Fidelity Polymerase displayed a notable increase in the percentage of chimeric reads starting from 25 cycles (Gohl et al. 2016) and had the second-highest rate of chimeras in comparison with other kits in 30 cycles (Sze and Schloss 2019). In this study, 25 cycles were used for all comparisons. Thus, the low performance, in terms of chimeras associated with those PCR kits, may result from cycle numbers and could potentially be reduced by reducing the number of cycles.

Previous studies investigating chimera formation and different polymerases report that reducing the amount of DNA template and the number of PCR cycles used (Oh et al., n.d.; Lahr and Katz 2009; Gohl et al. 2016), increasing the extension time (Qiu et al. 2001) and using high-fidelity (Oliver et al. 2015; Sze and Schloss 2019) or high-processivity polymerase (Gohl et al. 2016) can reduce the chimera formation. This is especially important for analysing environmental samples as chimeras appear more frequently in samples with high diversity (Fonseca et al. 2012). In addition, the unequal template concentrations in the environmental samples may contribute to the formation of chimeras (Lahr and Katz 2009). Several bioinformatics tools have been developed to identify and remove chimeras from the HTS data (Edgar et al. 2011; Haas et al. 2011; Quince et al. 2011; Wright et al. 2012). However, as the majority of chimeras may have more than two parent sequences, it might be difficult to detect the chimeras with the software (Lahr and Katz 2009). As an alternative approach, the usage of unique molecular identifiers has been demonstrated to be efficient for detecting chimeras and PCR-related bias in the HTS data (Filges et al. 2019; Fields et al. 2021).

The second parameter displaying a high number of significant differences between the PCR kits was BLAST top-hit similarity. The significant differences were mainly associated with four kits (KAPA2G, KAPA HiFi, Premix Ex Taq and Top DNA polymerase). Interestingly, from those PCR kits, Premix Ex Taq and KAPA2G also displayed lower results in some other parameters, for example, in terms of chimeras and base substitutions (KAPA2G) and in terms of chimeras, deletions and insertions (Premix Ex Taq). For one of the PCR kits (KAPA2G), the high proportion of errors may be explained by the lack of proofreading ability, whereas the second kit (Premix Ex Taq) has a proofreading ability. Thus, some other parameters or their combination might have a stronger influence on obtaining the correct sequences than the proofreading ability (Brandariz-Fontes et al. 2015).

The parameters associated with base substitutions also varied notably amongst the PCR kits. As those errors may also arise from sequencing (Glenn 2011; Pfeifer 2017), the results in those parameters may reflect a combination of PCR and sequencing-induced errors. Both the polymerase errors during PCR (substitutions) and sequencing errors (insertion, deletions, substitutions) have been reported as important sources of bias in the HTS data (Patin et al. 2013; Filges et al. 2019). A marginal contribution of polymerase-induced errors to the total errors in the HTS data has been demonstrated, based on the lack of statistically significant differences between the erroneous sequences obtained by proofreading and non-proofreading polymerases (Pfeiffer et al. 2018; Filges et al. 2019). In contrast, a clear pattern in the proportion of erroneous reads in the HTS data between the proofreading and non-proofreading polymerases has also been reported (Nichols et al. 2018). As the results of this study display statistically significant differences amongst different PCR kits and, as all the samples were pooled into one library analysed in a single run, all samples amplified with different PCR kits should experience similar sequencing errors (Lighten et al. 2014). Thus, the sequencing error might be responsible for some variability amongst the PCR kits, but the observed significant differences in terms of insertions, deletion and base substitutions are considered to be more influenced during PCR by the differences between PCR kits and the polymerase included.

Polymerase errors can be influenced by various parameters, such as template DNA, for example, complexity of the gene targeted (single allele or several alleles), presence of repetitive sequences, secondary structure (Cariello et al. 1991; Eckert and Kunkel 1991; Kunkel and Bebenek 2000; Brandariz-Fontes et al. 2015), the extent of DNA damage from extraction and thermocycling (Eckert and Kunkel 1991; Witzingerode et al. 1997; Potapov and Ong 2017) and by the PCR cycle number (Patin et al. 2013). The influence of the template DNA on the polymerase errors has been exemplified by the proportion of erroneous reads ranging from 8% to 50% between different polymerases when targeting a gene with a single allele (Brandariz-Fontes et al. 2015). This can partly be explained by the differences in the intrinsic error rate of each polymerase (Hestand et al. 2016). In general, a notably lower proportion of errors associated with insertions and deletions has been reported (<3%) to
be present in the HTS data compared to base substitution errors (> 96%; Kozich et al. 2013; Potapov and Ong 2017). Thus, the usage of high-fidelity polymerase could help reduce substitution errors originating from PCR as demonstrated in this study by two of the high-fidelity polymerases (Q5, KAPA HiFi). In addition, multi-template PCR, such as the metabarcoding approach using environmental DNA as a template, may be particularly susceptible to PCR biases due to the differences in template DNA sequences and their frequencies, leading to variations in amplification efficiencies (Kalle et al. 2014). Optimisation of PCR mixture component concentration and pH are known to improve the fidelity of polymerases (Ling et al. 1991). However, this should not be a notable influence as the PCR kits tested have been optimised by the manufacturers and, thus, the differences are assumed to reflect the intrinsic differences amongst PCR kits.

Interestingly, the amplification bias amongst species was generally low between the polymerases, except for Q5. The bias results from the high standard deviation between the species counts and should theoretically be similar for all the polymerases as the same mock community was analysed. Stochasticity of the PCR process has been shown to induce skewed sequence representation (Kebischull and Zador 2015). However, in this case, the bias should be greater between the PCR kits. The Q5 polymerase has previously demonstrated a low correlation between the observed and expected oligonucleotide proportions in metabarcoding ($R^2 = 0.44$; Nichols et al. 2018). Thus, the high amplification bias might be influenced by some intrinsic characteristics of this particular polymerase. Furthermore, the Q5 has been reported to prefer sequences with higher GC content (Nichols et al. 2018). However, as the GC content in the mock community varied in a narrow range (40% to 55%, an average of 44%, SD: 3%; data not shown), the differences in GC content are not expected to lead to the observed amplification bias. It has also been found that guanine-rich sequences can cause failure in PCR when proofreading polymerases are used (Zhu et al. 2016). As Q5 has one of the highest proofreading abilities compared to all the polymerases in the kits tested, it could serve as an explanation for the amplification bias amongst the species. However, in the mock community, the guanine content ranged from 24% to 29% (average: 26.8%, SD: 1%; data not shown) and, thus, it is not certain whether this difference is enough to result in high bias amongst the species. Further experiments are necessary to investigate the potential causes of amplification bias in Q5.

The majority of the kits tested in this study resulted in high values in terms of top-hit accuracy. This coincides with the results of a previous study finding no influence of polymerase choice on the species occurrence data in metabarcoding studies (Nichols et al. 2018). From all the kits analysed and conditions tested, two kits: KOD Plus NEO and HotStart Taq DNA Polymerase, both at 65 °C, displayed the highest values for the highest number of parameters associated with the top-hit similarity, deletions and chimeras. However, it is challenging to suggest the best PCR kit suitable for all metabarcoding studies as the differences in the diversity of the DNA template can also influence the outcome (Qiu et al. 2001). Thus, we recommend considering the importance of different parameters, for example, the seven parameters analysed, while planning a metabarcoding study and testing the suitability of the several PCR kits, for example, the three kits recommended here, for the particular samples.

Conclusions

The purpose of this study was to compare PCR kits containing polymerases to minimise PCR-based amplification artifacts in environmental DNA analysis, based on HTS and metabarcoding. Statistically significant differences amongst PCR kits were detected for all the parameters. The kits displaying significant variability as well as the best results for each parameter varied. The results of the comprehensive analysis visualised by radar charts suggested that KOD plus Neo and HotStart Taq DNA PCR kits with the annealing temperature of 65 °C displayed better performances in the highest number of parameters associated with chimeras, top hit similarity, and deletions. Especially, the higher annealing temperatures reduced chimaera formation. However, as the outcome may also be influenced by the template diversity (Qiu et al. 2001) and PCR cycling conditions (Patin et al. 2013), it is challenging to choose a single best PCR kit containing a polymerase that is suitable for all studies. Thus, we recommend using the knowledge generated in this study as a basis for choosing a PCR kit containing a polymerase in combination with further testing and optimisation for particular samples.

Acknowledgements

The authors thank R. Kubota for her assistance with the molecular work. This work was funded by “Technological developments for characterization of harmful plankton in the seawater”, Ministry of Agriculture, Forestry and Fisheries, Japan (16808839) [SN]; JST/JICA, Science and Technology Research Partnership for Sustainable Development (JPMJSA1705) [SN]; Japan Society for the Promotion of Science Short-term Postdoctoral Fellowship (PE18028) [SN, SS]; European Regional Development Fund and the programme Mobilitas Plus (MOBTP160) [SS]; Estonian Research Council grant (PSG735) [SS].

References


https://mbmg.pensoft.net


https://mbmg.pensoft.net
Satoshi Nagai et al.: Evaluation of DNA polymerases on PCR-generated artifacts


**Supplementary material 1**

**Table S1**

| Data type: accession numbers |
| Explanation note: **Table S1.** Information source and accession numbers on the nuclear ribosomal RNA gene (18S rRNA) of 40 species used for the mock sample. ND: no data. **Table S2.** Composition of PCR mixtures and PCR conditions. PCR amplification was done by using 14 PCR kits, for selected kits 3 different annealing temperatures were tested. **Table S3.** P-values resulting from posthoc comparisons of differences among PCR kits containing polymerase based on the seven parameters (A-G). Statistically significant p-values after Benjamini-Hochberg correction are marked with bold and in italics.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.6.77704.suppl1

**Supplementary material 2**

merge_tophit_count.pl

| Data type: Blast XML |

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.6.77704.suppl2
Supplementary material 3
same_tophit_count_merge.pl
Data type: Blast XML
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.6.77704.suppl3

Supplementary material 4
blastxml_parser
Data type: Blast XML
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/mbmg.6.77704.suppl4