Anal Bioanal Chem (2015) 407:1841–1848
DOI 10.1007/s00216-014-8435-y

RESEARCH PAPER

# Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples

**Marta Pawluczyk · Julia Weiss · Matthew G. Links · Mikel Egaña Aranguren · Mark D. Wilkinson · Marcos Egea-Cortines**

Received: 17 October 2014 / Revised: 15 December 2014 / Accepted: 18 December 2014 / Published online: 11 January 2015


**Abstract** Unbiased identification of organisms by PCR reactions using universal primers followed by DNA sequencing assumes positive amplification. We used six universal loci spanning 48 plant species and quantified the bias at each step of the identification process from end point PCR to next-generation sequencing. End point amplification was significantly different for single loci and between species. Quantitative PCR revealed that Cq threshold for various loci, even within a single DNA extraction, showed 2,000-fold differences in DNA quantity after amplification. Next-generation sequencing (NGS) experiments in nine species showed significant biases towards species and specific loci using adaptor-specific primers. NGS sequencing bias may be predicted to some extent by the Cq values of qPCR amplification.

M. Pawluczyk · J. Weiss · M. Egea-Cortines (✉)
Genetics, Instituto de Biotecnología Vegetal, Universidad Politécnica de Cartagena, 30202 Cartagena, Spain
e-mail: marcos.egea@upct.es

M. G. Links
Department of Computer Science, University of Saskatchewan, Saskatoon Research Centre, 107 Science Place, Saskatoon, SK S7N OX2, Canada

M. Egaña Aranguren · M. D. Wilkinson
Centro de Biotecnología y Genómica de Plantas UPM-INIA (CBGP), Campus Montegancedo, Autopista M-40 (Km 38), 28223 Pozuelo de Alarcón Madrid, Spain

M. Egaña Aranguren
Genomic Resources, Department of Genetics, Physical Anthropology and Animal Physiology, Faculty of Science and Technology, University of Basque Country (UPV/EHU), Sarriena auzoa z/g, 48940 Leioa-Bilbo, Spain

**Keywords** Metabarcoding · Next-generation sequencing · Ion torrent · Cq value · PCR efficiency

## Introduction

Sequence analysis of complex DNA samples is an important approach to monitoring species distribution in biodiversity and population studies. Genetic material is assessed using universal genomic sequences "barcodes" that are informative regarding the species composition of the sample, as they contain sufficient polymorphisms between species that taxonomic discrimination becomes possible [1]. The barcoding approach has become a mainstream technique to identify species in insects [2], very closely related plant species or hybrids [3], or fungi [4] and bacteria [5].

In plants, seven chloroplast *loci* have been analyzed as potential barcodes, the spacers *atpf-atph*, *trnH-psbA*, and *psbK-psbL* and the genes *matK*, *rbcL*, *rpoB*, and *rpoC1* [6, 7]. Metabarcoding involves DNA amplification of barcode loci from mixed-population samples, followed by next-generation sequencing (NGS). Sequenced fragments are then either assembled de novo and then aligned to known genome sequences [8] or are directly aligned to these genomic databases, thus becoming connected to specific taxa [9]. Most often, the objective of these analyses is to arrive at a quantitative measure of the relative abundance of the various species in the sample.

Despite being a proven tool for taxonomic identification, the approach of PCR is subject to a wide variety of potential biases throughout the processes of amplification and sequence

analysis, particularly when applied to mixed-population samples. These biases fall into three main categories. The first relates to differential barcode amplification success as a result of the barcode's universal primers. Depending on the marker/species combination, false-negative results can occur when sequence variation at the universal priming sites in one of the species prevents efficient annealing of the universal barcode primer for that species. A second type of bias relates to the efficiency of the amplification reaction, which may differ from species to species based on the sequence composition of their specific variant of the barcode. As a result, the proportion of sequences representing each species in the original sample may bear little resemblance to the proportion of that species in that population. Finally, there may also be biases introduced during the preparation of DNA libraries for sequencing. For instance, sample dilution has a strong effect on the correlation between biological and read quantities in bacterial samples [10]. A combination of barcoding and NGS has been in some cases confirmed by qPCR, showing that while the exact quantification is not precise, trends in the population structure are faithful [11].

Despite knowing that these potential biases exist, the degree to which each source of bias affects the outcome of a metabarcoding experiment and their relative importance have not been well quantified. Moreover, by quantifying these biases and relating them to the specific sequences being studied, it may be possible to formulate approaches for post facto normalization of metabarcode data to better reflect the population makeup. For example, PCR efficiency is an important parameter of quantitative PCR analysis of gene expression [12–14], and while a variety of algorithms exist that predict the efficiency of PCR amplification, these are currently not considered in any of the normal barcoding or metabarcoding pipelines. Amplification efficiency for a given DNA sequence depends heavily on the G + C content of the amplicon [14], DNA secondary structure [15], and previous sample treatment [16]. Under optimal PCR conditions with 100 % amplification efficiency, two copies of DNA are generated from each template during exponential phase of amplification, and such a reaction is said to have an efficiency of 2. This efficiency can also affect another important statistic, namely Cq a relative measure of the predicted concentration of the target amplicon in a PCR reaction and a measurement that is widely used in qPCR analysis [17, 18]. These kinds of statistics will be even more relevant to NGS technologies that introduce additional PCR amplification steps, such as Ion Torrent or 454/Roche that utilizes an emulsion PCR during library construction [19].

The present study, therefore, aims to first quantitatively analyze PCR success and evaluate amplification efficiency and Cq values as a tool for predicting amplification success. In this study, we undertake a survey of six well-known plant barcoding markers and apply them to 48 species from 34 different plant families. In addition, we apply the Ion Torrent sequencing method simultaneously for mixed-species PCR products of three barcoding primers *rbcL*, *rpoB*, and *rpoC1* starting with equal amounts of PCR products, to quantitatively measure the bias introduced by this step of the metabarcoding study.

Our results reveal that quantitative and even qualitative interpretation of metabarcoding data based on read abundance is fraught with potential, serious biases. We present, in detail, a dissection of the degree of bias introduced at each step in the typical laboratory practice of barcode marker analysis from mixed DNA samples.

## Materials and methods

### Plant material

Plant material, 48 plant species belonging to 33 different families, was gathered from the local fruit market, field sampling, botanical records, and our own collections (Table 1).

### DNA extraction and real-time PCR

Two independent genomic DNA samples were extracted from fresh leaf using the commercial kit "Plant NucleoSpin" (Machery and Nagel, Düren, Germany). All extracted samples were quantified with a Nanodrop 2000 and, after isopropanol-ethanol precipitation, all samples were diluted to 50 ng/μl in order to have identical concentrations. Single species reactions were performed from the two independent DNA extractions with three technical replicas for a total of six PCR reactions per species using 100-ng DNA/reaction. Real-time PCR reactions were performed as described previously [14]. The primers used in this experiment (*rbcL*-a, *matK*, *rpoB*, *rpoC1*, *trnL-F*, *trnH-psbA*) have been described previously [6].

Equal amounts of genomic DNA from three species were used to create the mixed-species metabarcoding templates. Amplifications were performed using an initial DNA quantity of 150 ng corresponding to 50 ng of each of the three genomes. Sequencing reactions comprised nine species.

### qPCR efficiency and Cq calculation

qPCR efficiency and Cq were computed using *qpcR*, R package [20]. Efficiency value ($E$) was calculated as $E_{cpD2} = F(cpD2)/F(cpD2) - 1$, in which $F$ is raw fluorescence at cycle $x$, and cpD2 is cycle number at second derivative maximum of the curve [21].

**Table 1** List of plant species analyzed

| Plant species | Family | Location/donor population |
|---|---|---|
| *Spinacia oleracea* | Amaranthaceae | Murcia, Spain/commercial |
| *Pistacia lentiscus* | Anacardiaceae | Murcia, Spain/natural |
| *Daucus carota* | Apiaceae | Murcia, Spain/commercial |
| *Nerium oleander* | Apocynaceae | Murcia, Spain/artificial |
| *Arisarum vulgare* | Araceae | Murcia, Spain/natural |
| *Phoenix dactylifera* | Arecaceae | Murcia, Spain/commercial |
| *Aloe vera* | Asphodelaceae | Murcia, Spain/artificial |
| *Lactuca sativa* | Asteraceae | Murcia, Spain/commercial |
| *Cynara scolymus* | Asteraceae | Murcia, Spain/commercial |
| *Brassica oleracea botrytis* | Brassicaceae | Murcia, Spain/commercial |
| *Brassica oleracea italica* | Brassicaceae | Murcia, Spain/commercial |
| *Diplotaxis erucoides* | Brassicaceae | Murcia, Spain/natural |
| *Lobularia maritima* | Brassicaceae | Murcia, Spain/natural |
| *Arabidopsis thaliana* | Brassicaceae | Murcia, Spain/artificial |
| *Silene vulgaris* | Caryophyllaceae | Murcia, Spain/natural |
| *Cistus albidus* | Cistaceae | Murcia, Spain/natural |
| *Cistus heterophyllus* | Cistaceae | Murcia, Spain/natural |
| *Aeonium arboreum* | Crassulaceae | Murcia, Spain/natural |
| *Cucumis sativus* | Cucurbitaceae | Biala Podlaska, Poland/ commercial |
| *Ecballium elaterium* | Cucurbitaceae | Murcia, Spain/natural |
| *Chamaecyparis* sp. | Cupressaceae | Murcia, Spain/artificial |
| *Arbutus unedo* | Ericaceae | Murcia, Spain/artificial |
| *Ricinus communis* | Euphorbiaceae | Murcia, Spain/artificial |
| *Ceratonia siliqua* | Fabaceae | Murcia, Spain/natural |
| *Pisum sativum* | Fabaceae | Murcia, Spain/artificial |
| *Vicia faba* | Fabaceae | Murcia, Spain/artificial |
| *Quercus coccifera* | Fagaceae | Murcia, Spain/natural |
| *Pelargonium* × *hortorum* | Geraniaceae | Murcia, Spain/artificial |
| *Leucobryum glaucum* | Leucobryaceae | Biala Podlaska, Poland/ natural |
| *Anagallis arvensis* | Myrsinaceae | Murcia, Spain/natural |
| *Callistemos* sp. | Myrtaceae | Murcia, Spain/artificial |
| *Olea europaea* | Oleaceae | Murcia, Spain/artificial |
| *Oxalis pes-caprae* | Oxalidaceae | Murcia, Spain/natural |
| *Pinus silvestres* | Pinaceae | Biala Podlaska, Poland/ natural |
| *Antirrhinum majus* | Plantaginaceae | Murcia, Spain/artificial |
| *Zea mays* | Poaceae | Murcia, Spain/commercial |
| *Oryza sativa* | Poaceae | Murcia, Spain/artificial |
| *Hordeum vulgare* | Poaceae | Murcia, Spain/commercial |
| *Piptatherum miliaceum* | Poaceae | Murcia, Spain/natural |
| *Portulacaria afra* | Portulacaceae | Murcia, Spain/artificial |
| *Galium verrucosum* | Rubiaceae | Murcia, Spain/natural |
| *Populus alba* | Salicaceae | Murcia, Spain/artificial |
| *Petunia hybrida* | Solanaceae | Murcia, Spain/artificial |
| *Solanum tuberosum* | Solenaceae | Murcia, Spain/commercial |
| *Solanum lycopersicum* | Solenaceae | Murcia, Spain/commercial |
| *Thymelaea hirsuta* | Thymelaeaceae | Murcia, Spain/natural |
| *Vitis vinifera* | Vitaceae | Murcia, Spain/commercial |
| *Asphodelus fistulosus* | Xanthorrhoeaceae | Murcia, Spain/natural |

### Determination of relative abundance of sequences from PCR products of mixed genomic DNA by semiconductor sequencing

PCR products generated by amplifying, separately, the chloroplast barcoding sequences *rbcL*-a, *rpoC1*, and *rpoB* from mixed genomic DNAs (100 ng each) were pooled equivalently to yield a final amount of 100 ng. Initial time of digestion was adjusted to yield 300-bp fragments. Preparation of samples for library construction and sequencing were performed using the Ion Torrent Next-Generation Sequencing Kits (Life Technologies, CA, USA) according to the manufacturer's instructions. Briefly, PCR products were fragmented using the Ion Shear Plus reagent to a fragment size of 200 bp. The corresponding fragments were ligated to adaptors and size fractionated using E-Gel electrophoresis, obtaining fragments of average 330 bp. Emulsion PCR was performed using one-touch system according to the manufacturer's protocol, and sequencing was performed using 314 Ion Torrent chips. A total of 333,274 reads with a mean read length of 159 bp were computationally analyzed in order to identify species origin of each fragment by aligning the reads with a library of known chloroplast sequences using Bowtie2 [22]. We extracted from the resulting SAM file a map of reads to the known chloroplast sequences using a Perl script from the mPuma pipeline [8]. The analysis can be reproduced, with the same parameters and data, at the following Galaxy installation (page: http://biordf.org:8983/u/mikel-egana-aranguren/p/sources-of-bias-in-applying-barcoding-markers-for-sequence-analysis-of-environmental-samples).

## Results

This work aimed to reveal and quantify the biases that can occur during metabarcoding analyses. We executed our analyses using the most widely accepted plant barcodes, quantitated our results using widely accepted practices such as qPCR, and followed normal protocols for library construction and NGS. At each stage, we re-normalized the samples such that we knew the precise quantities and relative abundances of the input DNA. In addition, although it is known that the size

of the PCR amplification product plays a major role in bias within bacterial community pyrosequencing projects [23], the size of the amplicons analyzed here is below the 1-Kb threshold identified in those studies. Thus, we should be able to safely exclude that as a possible cause of bias in this study.

## Suitability of barcodes depending on plant species

The worst possible outcome of a metabarcode analysis is false-negative, i.e., lack of amplification of a species barcode despite presence of that taxon in the population. As such, our first analysis assessed PCR success. As expected, it varied both between barcode markers and between the 48 plant species tested. Barcode primers for the *matK* gene were the least successful, giving positive results in only 50 % of the tested species, followed by *rbcL* which amplified in 82 % of species. The *rpoB and rpoC1* genes as well as the short intergenic spacers *trnL-F* and *trnH-psbA* proved to be the most universally successful barcoding markers, amplifying in close to 90 % of the investigated species. Our data, however, gives a within species assessment of PCR success based on six independent amplifications. As none of the samples had a complete failure of amplification with all primer combinations, we can conclude that DNA quality was not a limiting factor for amplification.

## qPCR parameters for specific barcodes depending on plant species

The second phase of the analysis addressed whether end point PCR results are the outcome of PCR efficiency. As shown in Fig. 1, amplification efficiency during qPCR varied between barcode markers. The highest average efficiency, based on amplification from all species, corresponded to the markers *trnL-F* and *trnH-psbA* followed by *rpoB*, *rpoC1*, and *rbcL*. The *matK* barcode showed the lowest average efficiency among all species. The efficiencies of *matK*, *rbcL*, and *rpoC1*, but not *rpoB* and *trnH-psbA*, were significantly

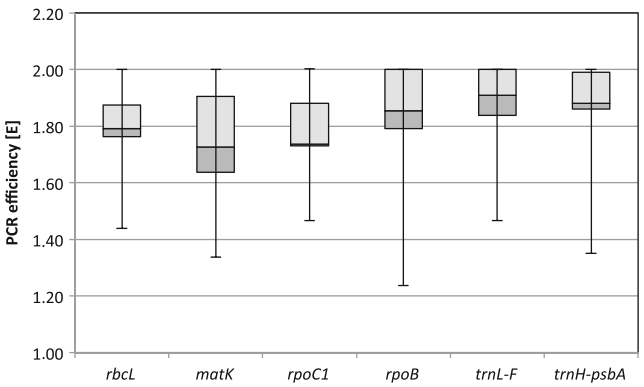


**Fig. 1** Boxplot of PCR efficiency data for six barcoding markers derived from qPCRs of 48 plant species. The graphic shows only successful amplification data with an efficiency >1

different from high-efficiency marker *trnL-F* ($p<0.0001$ for *matK* and *rbc*L and $p=0.0013$ for *rpoC1*). PCR efficiencies considering all barcode markers for selected species are summarized in Table 2 showing that both the barcode target and the species are amplified from govern efficiency.

As PCR success could be the result of initial priming and some samples gave no amplification, we compared the priming site for the worst performing pair of primers (2.1.f matK and 5r matk) with their corresponding priming sites of negative performers *Zea mays*, *Quercus coccifera*, and *Brassica oleracea*, *Oryza sativa* as middle quality, and *Vitis vinifera* that had the best overall amplification with this marker (Fig. 2). Indeed, mispriming may explain the lack of amplification in the case of *Z. mays*, but it is not obvious the differences in the other samples. Furthermore, amplification efficiency may be affected by other parameters beyond priming (see below).

Looking at intra-species variation for all barcodes, Cq values varied widely in this case also (Fig. 3 and Table 3). Some extreme cases of intraspecific variation were found in *Oryza sativa* where *rbcL* showed no amplification, whereas *trnL-F* had a Cq of 11.93 (Table 3). Beyond the false-negatives, other important differences in Cq were observed for the various markers. In *O. sativa*, the difference in Cq between *matK* (28.55) and *trnL-F* (11.93) is extremely large. If one were to apply the delta-CT formula [18], and assumed an average efficiency for both markers (efficiency=1.9), the predicted differences in starting DNA level would be 2,116-fold based on the estimates from these two barcodes. This was not an isolated case as we found negative amplification of *rbcL* or *matK* and positive albeit differing Cq values in 20 % of the species tested for this parameter (*Z. mays*, *Daucus carota*, *Q. coccifera*, and *Asphodelus fistulosa*).

Cq values also varied significantly among species considering all six markers together, and these differences did not correlate with the average efficiency of the PCR amplification. For example, *Z. mays* exhibited an average efficiency over all barcodes of 1.88±0.08 and an average Cq of 30.76±4.67, while *Solanum tuberosum* exhibited a similar average efficiency of 1.86±0.15, yet had a Cq of 15.98±5.30. Moreover, for any given barcode, PCR efficiency and Cq values also proved to be independent variables, based on regression analysis ($R^2$ between 0.37 and 0.003).

Differences in efficiency or Cq may be related to amplification bias among template DNAs in environmental samples. We analyzed abundance of reads after sequencing in order to address this question.

## Biases during pre-amplification and during emulsion PCR

The identification of genomic DNAs corresponding to different organisms in environmental samples requires sequencing of barcode-PCR products. Not all barcodes successfully

**Table 2** PCR efficiency evaluated in a selection of plant species

| Plant family | *rbcL-a* | *matK* | *rpoC1* | *rpoB* | *trnL-F* | *trnH-psbA* | Average±SD |
|---|---|---|---|---|---|---|---|
| Oxalidaceae (*Oxalis pes-caprae*) | 1.89 | 1.83 | 1.70 | 1.78 | 1.91 | 1.90 | 1.84±0.08 |
| Cistaceae (*Cistus heterophyllus*) | 1.83 | 1.80 | 1.66 | 1.71 | 1.90 | 1.95 | 1.81±0.11 |
| Poaceae (*Zea mays*) | 1.85 | NA | 1.72 | 1.97 | 1.80 | 1.91 | 1.85±0.10 |
| Oleaceae (*Olea europaea*) | 1.76 | 1.51 | 1.79 | 1.88 | 1.93 | 1.95 | 1.80±0.16 |
| Salicaceae (*Populus alba*) | 1.78 | 1.78 | 1.78 | 1.89 | 1.98 | 1.98 | 1.87±0.10 |
| Poaceae (*Oryza sativa*) | NA | 1.82 | 1.79 | 1.72 | 1.98 | 1.81 | 1.82±0.10 |
| Apiaceae (*Daucus carota*) | 1.94 | NA | 1.85 | 2.00 | 1.98 | 2.00 | 1.95±0.06 |
| Solananceae (*Solanum tuberosum*) | 1.70 | 1.70 | 1.85 | 1.84 | 1.95 | 2.00 | 1.80±0.12 |
| Scrophulariaceae (*Antirrhinum majus*) | 1.79 | 1.82 | 1.98 | 1.99 | 2.00 | 2.00 | 1.93±0.1 |
| Arecaceae (*Phoenix dactylifera*) | 1.87 | 1.90 | 1.97 | 1.97 | 2.00 | 1.84 | 1.92±0.06 |
| Cucurbitaceae (*Cucumis sativus*) | 1.84 | 1.80 | 1.91 | 1.99 | 1.98 | 1.91 | 1.9±0.07 |
| Amaranthaceae (*Spinacia oleracea*) | 1.90 | 1.42 | 1.99 | 2.00 | 2.00 | 1.99 | 1.88±0.23 |
| Vitales (*Vitis vinifera*) | 1.82 | 1.85 | 1.75 | 1.94 | 1.89 | 1.95 | 1.87±0.08 |
| Solanaceae (*Petunia hybrida*) | 1.73 | 1.73 | 1.86 | 1.85 | 1.93 | 1.94 | 1.84±0.09 |
| Fabaceae (*Ceratonia silique*) | 1.83 | 1.70 | 1.84 | 1.79 | 1.91 | 1.91 | 1.83±0.08 |
| Fagaceae (*Quercus coccifera*) | NA | NA | 1.68 | 1.72 | 1.90 | 1.86 | 1.79±0.11 |
| Thymelaeaceae (*Thymelea hirsuta*) | 1.88 | NA | 1.73 | 1.78 | 1.81 | 1.75 | 1.79±0.06 |
| Xanthorrhoeaceae (*Asphodelus fistulosus*) | 1.81 | NA | 1.73 | 1.76 | 1.78 | 1.84 | 1.78±0.04 |
| Brasicaceae (*Brassica oleracea*) | 1.70 | NA | 1.76 | 1.82 | 1.76 | 1.67 | 1.74±0.06 |
| Asteraceae (*Cynara Scolymus*) | 1.49 | 1.62 | 1.50 | 1.49 | 1.49 | 1.40 | 1.5±0.07 |
| Average | 1.80 | 1.73 | 1.79 | 1.84 | 1.89 | 1.88 | |
| Standard deviation | 0.10 | 0.14 | 0.12 | 0.13 | 0.12 | 0.14 | |

Samples with NA were non-successful PCR amplifications

amplify in each species. Table 4 shows the result of simultaneous sequencing of equal amounts of PCR products from mixed-species templates amplified with barcode markers, *rbcL*, *rpoB*, and *rpoC1*. The results reveal a strong bias in the number of reads corresponding each species contained in the equimolar starting sample. In the case of marker *rpoB*, most reads (95 %) corresponded to *S. tuberosum* and only 0.02 % to *Z. mays.* The number of reads was not related to the PCR efficiencies of the species but was related to their Cq values when amplified separately (Table 4).

Analysis of read numbers also showed a strong bias in the number of total reads corresponding to each of the barcodes (Table 4). Although equal amounts of PCR product from pre-amplification were used to create the amplicon library, only 11.2 % of all reads were identified as *rbcL* fragments, 36.5 % as *rpoB* fragments, and 52.3 % as *rpoC1* fragments. These

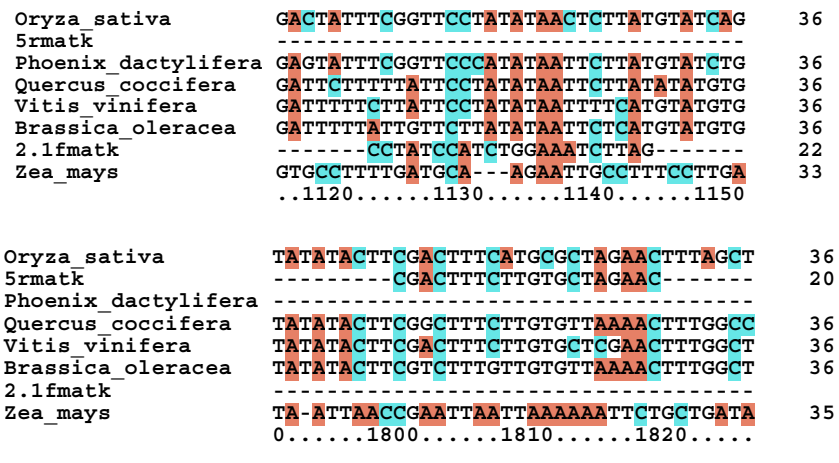


**Fig. 2** Annealing of primers 2.1f-matk and 5rmatk to sequences rendering negative amplification (*Quercus coccifera*, *Brassica oleracea*, and *Zea mays*) and positive amplification (*Oryza sativa*, *Vitis vinifera*, and *Phoenix dactylifera*)

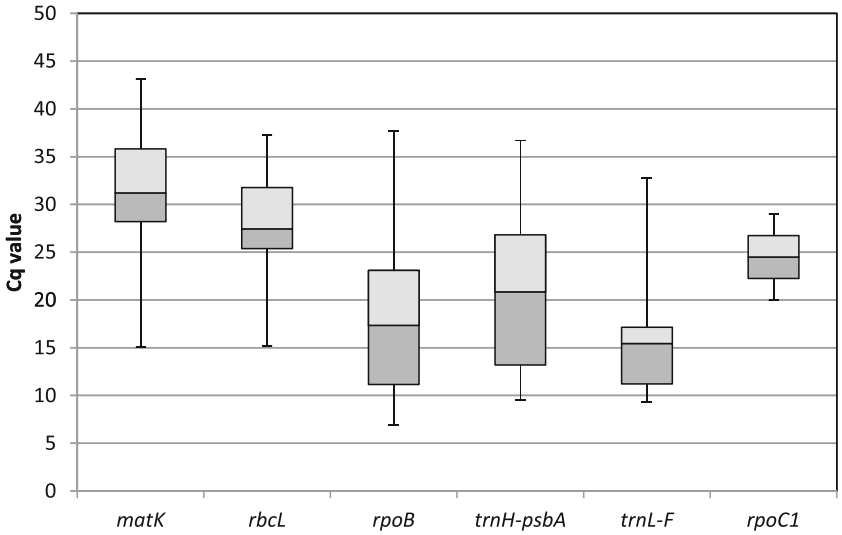


**Fig. 3** Boxplot of Cq values for six barcoding markers derived from qPCRs of 48 plant species

results are significantly different from an expected 33.3 % per reaction (chi-square test $p<2.2$ e-16). The relative percentages in read number proved independent of PCR efficiencies of the specific markers but correlated with average Cq values of the marker for the three species amplified.

As emulsion PCR for NGS sequencing is performed with primers that correspond to ligated adaptors, and nevertheless a relationship between Cq values and final number of reads is maintained, we can conclude that the main bias that can be encountered in metabarcoding projects is related to the specific sequence of the barcode fragment. This seems to be independent of any primer-specific effect such as internal priming, etc., as it is consistent over two different primer pairs. Library construction can produce at least 4.6-fold differences when comparing *rbcL* against *rpoC1*.

**Table 3** Cq qPCR values obtained in a selection of plant species

| Plant family | *rbcL-a* | *matK* | *rpoC1* | *rpoB* | *trnL-F* | *trnH-psbA* | Average±SD |
|---|---|---|---|---|---|---|---|
| Oxalidaceae (*Oxalis pes-caprae*) | 30.99 | 36.24 | 22.63 | 23.44 | 19.41 | 27.76 | 26.75±6.18 |
| Cistaceae (*Cistus heterophyllus*) | 25.83 | 28.80 | 24.85 | 25.01 | 16.74 | 18.86 | 23.35±4.58 |
| Poaceae (*Zea mays*) | 34.74 | NA | 22.35 | 25.17 | 20.15 | 26.06 | 25.69±5.57 |
| Oleaceae (*Olea europaea*) | 26.05 | 23.86 | 17.82 | 15.18 | 16.74 | 17.52 | 19.53±4.36 |
| Salicaceae (*Populus alba*) | 24.13 | 29.89 | 15.29 | 13.82 | 13.25 | 13.90 | 18.38±6.96 |
| Poaceae (*Oryza sativa*) | NA | 28.55 | 14.52 | 22.77 | 11.93 | 25.02 | 20.56±7.06 |
| Apiaceae (*Daucas carota*) | 15.82 | NA | 13.06 | 9.77 | 20.15 | 25.95 | 26.95±6.31 |
| Solananceae (*Solanum tuberosum*) | 16.77 | 20.55 | 10.16 | 8.65 | 10.53 | 10.90 | 12.93±4.66 |
| Scrophulariaceae (*Antirrhinum majus*) | 27.81 | 33.83 | 13.06 | 12.72 | 12.06 | 15.08 | 19.09±9.34 |
| Arecaceae (*Phoenix dactylifera*) | 31.39 | 16.06 | 10.81 | 15.32 | 10.12 | 19.95 | 17.28±7.81 |
| Cucurbitaceae (*Cucumis sativus*) | 27.17 | 29.71 | 9.89 | 9.13 | 9.02 | 23.57 | 18.08±9.77 |
| Amaranthaceae (*Spinacia oleracea*) | 29.66 | 19.59 | 8.94 | 25.32 | 9.40 | 10.40 | 17.22±8.97 |
| Vitales (*Vitis vinifera*) | 33.15 | 18.17 | 17.65 | 13.66 | 13.88 | 15.48 | 18.67±7.34 |
| Solanaceae (*Petunia hybrida*) | 28.38 | 19.47 | 11.02 | 10.28 | 10.42 | 11.03 | 15.10±7.40 |
| Fabaceae (*Ceratonia silique*) | 32.84 | 23.26 | 16.13 | 18.73 | 14.99 | 20.09 | 21.01±6.50 |
| Fagaceae (*Quercus coccifera*) | NA | NA | 23.39 | 18.43 | 17.06 | 25.14 | 21.01±3.87 |
| Thymelaeaceae (*Thymelea hirsuta*) | 29.52 | NA | 14.70 | 24.30 | 16.52 | 27.4 | 22.49±6.58 |
| Xanthorrhoeaceae (*Asphodelus fistulosus*) | 26.73 | NA | 19.38 | 18.13 | 18.91 | 22.84 | 21.20±3.58 |
| Brasicaceae (*Brassica oleracea*) | 24.55 | NA | 14.76 | 13.57 | 14.35 | 21.83 | 17.81±5.02 |
| Asteraceae (*Cynara Scolymus*) | 34.47 | 32.27 | 23.89 | 23.45 | 23.27 | 22.94 | 26.72±5.21 |
| Average | 27.78 | 25.73 | 16.22 | 17.34 | 14.95 | 20.09 | |
| Standard deviation | 5.28 | 6.41 | 5.09 | 5.90 | 4.13 | 5.69 | |

Samples with NA correspond to unsuccessful amplifications

**Table 4** Average PCR efficiencies ($PCR_{eff}$), Cq values, and sequence reads derived from PCR products of barcodes *rbcL*, *rpoB*, and *rpoC1* using ion semiconductor sequencing

| | Barcoding locus | | | | |
|---|---|---|---|---|---|
| | *rbcL* | | % of reads | $PCR_{eff}$ of the species | Cq of the species |
| Average $PCR_{eff}$ for the amplified species (together) | 1.81±0.09 | *Oxalis pes-caprae* | 0.87 | 1.89±0.04 | 30.99±0.82 |
| Average Cq for the amplified species (together) | 26.97±7.52 | *Vitis vinifera* | 4.21 | 1.82±0.02 | 33.15±0.78 |
| Total reads | 34,239 | *Solanum tuberosum* | 94.92 | 1.69±0.04 | 16.77±0.88 |
| % of total reads | 11.2 | | | | |
| | *rpoB* | | | | |
| Average $PCR_{eff}$ for the amplified species (together) | 1.85±0.14 | *Zea mays* | 0.02 | 1.71±0.13 | 25.01±0.7 |
| Average Cq for the amplified species (together) | 21.79±5.00 | *Cistus heterophyllus* | 1.13 | 1.97±0.06 | 25.17±0.27 |
| Total reads | 111,407 | *Olea europaea* | 98.85 | 1.86±0.01 | 16.28±0.26 |
| % of total reads | 36.5 | | | | |
| | *rpoC1* | | | | |
| Average $PCR_{eff}$ for the amplified species (together) | 1.74±0.06 | *Cistus heterophyllus* | 0.34 | 1.66±0.04 | 24.85±1.24 |
| Average Cq for the amplified species (together) | 18.22±4.96 | *Oryza sativa* | 36.57 | 1.79±0.02 | 14.52±0.54 |
| Total reads | 159,923 | *Populus alba* | 63.09 | 1.78±0.03 | 15.29±1.51 |
| % of total reads | 52.3 | | | | |

## Discussion

Similarities between primer and template, as well as the regional G + C content of a template, are factors that influence PCR efficiency [22, 24]. The low PCR success, particularly in case of *matK* with 50 % PCR failure in a screening of 48 species, is probably due to lack of similarity between primer and template, since no highly conserved sites flanking the most variable parts of this barcoding marker exist [7]. Indeed, indels and mispriming may account for lack of success in PCR amplification (see Fig. 2). However, it is not a straightforward assessment to understand the lack of amplification that may be also the result of specific features of the DNA strand amplified.

The Cq parameter is widely used in qPCR analysis [17, 18], and we applied this to assess intraspecific and interspecific variability in both PCR success and as a possible parameter to estimate final read numbers in NGS experiments. Surprisingly, there was a wide range of Cq values identified within a single species, and even within a single DNA extraction, something completely unexpected as Cq values are thought to relate to DNA/cDNA quantities. These ranges were far beyond the 1–2 cycles that might arise from sampling and manipulation errors.

Our results show that PCR efficiency varies among barcoding markers and species but that these differences in efficiency do not relate to the corresponding Cq values as measure of PCR success. The Cq values in contrast proved to be a valuable parameter for the estimation of PCR success as *matK* and *rbcL* showed the highest Cq values during qPCR. The late take-off in the qPCR assay for *rbcL* and *matK* probably reflect an excess of mismatches between primers and templates as Cq values also varied significantly among species over the whole range of markers that may be related to DNA quality and/or PCR inhibiting substances contained in the sample.

One of the most common aims in analyzing environmental samples is to estimate the relative abundance of species based on determining the quantity of their template DNAs. In principle, equal amounts of template DNA from different species should lead to 1:1 amplicon numbers. However, Suzuki and Giovannoni (1996) observed preferential amplification of certain bacterial fragments in mixed templates with lower G + C content [23]. Our results show the situation is similar in plants, with a strong bias in relative read number among three species after Ion Torrent sequencing. Low read numbers corresponded to species with high Cq values for a given marker, whereas PCR efficiency seemed unrelated, indicating that species with lower Cqs for a given marker are preferentially amplified.

As such, further improving the reliability of amplification and utilization of sequence content features to derive and apply quantitative data normalization algorithms are certainly areas of significant interest for future development in metabarcoding and NGS analysis.

**Acknowledgments** This work was performed as partial fulfillment of the PhD of Marta Pawluczyk. This work was funded by the Comunidad Autónoma de la Región de Murcia Project "Molecular markers in conservation and management of the flora of Murcia Region" ("Marcadores moleculares en conservación y gestión de la flora murciana"). Part of the work was performed under the Proyecto Vitalis Campus Mare Nostrum

"Espacio Mediterráneo de Investigación en Red en Alimentos y Salud" - CEI10-2-0002.

**Data availability** Raw and processed data will be made publicly available via entries in Data Dryad, and a formal Data Descriptor will be published detailing the methodologies and workflows used, as well as rich descriptions of the data elements themselves. The analytical workflow for sequence processing and mapping is already publicly available as a Galaxy workflow, as described in the manuscript, and can be freely re-run at any time. The analysis can be reproduced, with the same parameters and data, at the following Galaxy installation (page: http://biordf.org:8983/u/mikel-egana-aranguren/p/sources-of-bias-in-applying-barcoding-markers-for-sequence-analysis-of-environmental-samples).

## References


1. Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. Trends Genet 23:167–172. doi:10.1016/J.Tig.2007.02.001
2. Hajibabaei M, Janzen DH, Burns JM et al (2006) DNA barcodes distinguish species of tropical Lepidoptera. Proc Natl Acad Sci U S A 103:968–971
3. Pawluczyk M, Weiss J, Vicente-Colomer MJ, Egea-Cortines M (2012) Two alleles of rpoB and rpoC1 distinguish an endemic European population from Cistus heterophyllus and its putative hybrid (C. x clausonis) with C. albidus. Plant Syst Evol 298:409–419
4. Krüger M, Stockinger H, Krüger C et al (2009) DNA-based species level detection of Glomeromycota: one PCR primer set for all arbuscular mycorrhizal fungi. New Phytol 183:212–223. doi:10.1111/j.1469-8137.2009.02835.x
5. Links MG, Dumonceaux TJ, Hemmingsen SM, Hill JE (2012) The chaperonin-60 universal target is a barcode for bacteria that enables de novo assembly of metagenomic sequence data. PLoS ONE 7:e49755. doi:10.1371/journal.pone.0049755
6. Hollingsworth PM, Forrest LL, Spouge JL et al (2009) A DNA barcode for land plants. Proc Natl Acad Sci U S A 106:12794–12797. doi:10.1073/Pnas.0905845106
7. Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region. PLoS ONE 2:e508. doi:10.1371/journal.pone.0000508
8. Links MG, Chaban B, Hemmingsen SM et al (2013) mPUMA: a computational approach to microbiota analysis by de novo assembly of operational taxonomic units based on protein-coding barcode sequences. Microbiome 1:23. doi:10.1186/2049-2618-1-23
9. Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. Mol Ecol 21:1834–1847. doi:10.1111/j.1365-294X.2012.05550.x
10. Amend AS, Seifert KA, Bruns TD (2010) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? Mol Ecol 19:5555–5565. doi:10.1111/j.1365-294X.2010.04898.x
11. Links MG, Demeke T, Gräfenhan T et al (2014) Simultaneous profiling of seed-associated bacteria and fungi reveals antagonistic interactions between microorganisms within a shared epiphytic microbiome on Triticum and Brassica seeds. New Phytol. doi:10.1111/nph.12693
12. Platts AE, Johnson GD, Linnemann AK, Krawetz SA (2008) Real-time PCR quantification using a variable reaction efficiency model. Anal Biochem 380:315–322
13. Pfaffl MW, Horgan GW, Dempfle L (2002) Relative expression software tool (REST(C)) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Res 30:e36. doi:10.1093/nar/30.9.e36
14. Mallona I, Weiss J, Egea-Cortines M (2011) pcrEfficiency: a Web tool for PCR amplification efficiency prediction. BMC Bioinforma 12:404. doi:10.1186/1471-2105-12-404
15. D'haene B, Vandesompele J, Hellemans J (2010) Accurate and objective copy number profiling using real-time quantitative PCR. Methods 50:262–270. doi:10.1016/j.ymeth.2009.12.007
16. Von Holst C, Boix A, Marien A, Prado M (2012) Factors influencing the accuracy of measurements with real-time PCR: the example of the determination of processed animal proteins. Food Control 24:142–147. doi:10.1016/j.foodcont.2011.09.017
17. Bustin SA, Benes V, Garson JA et al (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem 55:611–622. doi:10.1373/clinchem.2008.112797
18. Schmittgen TD, Livak KJ (2008) Analyzing real-time PCR data by the comparative CT method. Nat Protoc 3:1101–1108. doi:10.1038/nprot.2008.73
19. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141. doi:10.1016/j.tig.2007.12.007
20. Ritz C, Spiess AN (2008) qpcR: an R package for sigmoidal model selection in quantitative real-time polymerase chain reaction analysis. Bioinformatics 24:1549–1551. doi:10.1093/Bioinformatics/Btn227
21. Spiess A-N, Feig C, Ritz C (2008) Highly accurate sigmoidal fitting of real-time PCR data by introducing a parameter for asymmetry. BMC Bioinforma 9:221. doi:10.1186/1471-2105-9-221
22. Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. Appl Environ Microbiol 64:3724–3730
23. Suzuki M, Giovannoni S (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl Environ Microbiol 62:625–630
24. Benita Y, Oosting RS, Lok MC, et al. (2003) Regionalized GC content of template DNA as a predictor of PCR success. 31:1–7. doi: 10.1093/nar/gng101