Quantitative evaluation of bias in PCR amplification and Next Generation Sequencing derived from metabarcoding samples

M. Pawluczyk[(1)], J.Weiss[(1)], M.G. Links[(2)], M.E. Aranguren[(3,4)], M.D. Wilkinson[(3)], M. Egea-Cortines[(1)]

[(1)] Dirección Genetics, Instituto de Biotecnología Vegetal, Universidad Politécnica de Cartagena, 30202, Cartagena, Spain, e-mail: marta.pawluczyk@gmail.com
[(2)] Department of Computer Science, University of Saskatchewan, Saskatoon Research Centre, 107 Science Place Saskatoon, SK, S7N OX2, Canada
[(3)] Centro de Biotecnología y Genómica de Plantas UPM-INIA (CBGP), Campus Montegancedo, Autopista M-40 (Km 38), 28223-Pozuelo de Alarcón Madrid, Spain
[(4)] 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

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 2000-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 Q-PCR amplification.

Keywords: Q-PCR, Ion torrent; Cq value; PCR efficiency

1. Introduction

Sequence analysis of complex DNA samples, such important in population and biodiversity studies, base on universal genomic sequences “barcodes”. These sequences are informative regarding the species composition of the sample, as they contain sufficient polymorphisms between species that taxonomic discrimination becomes possible [1]. This approach constitutes a mainstream technique in species identification. 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, rpoC1* [2,3]. Metabarcoding involves DNA amplification of barcode loci from mixed samples, followed by Next-Generation Sequencing (NGS) in order to identify specific taxa. 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.

Unfortunately this widely applied tool is subject to a wide variety of potential biases that can be grouped in three categories related to: 1 – the differential barcode amplification success as a result of the barcode’s universal primers, 2 - 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, 3 - the preparation of DNA libraries for sequencing.

Despite knowing that these potential biases exist, the degree to which each source of bias affects the outcome of a metabarcoding experiment 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 make-up.

The present study, therefore, aims to first quantitatively analyze PCR success and evaluate amplification efficiency and Cq values (a relative measure of the predicted concentration of the target amplicon in a PCR reaction [4]) 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 interpretation of metabarcoding data based on read-abundance is fraught with potential, serious biases.

2. Materials and Methods

2.1 Plant material

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

2.2 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, Germany). All extracted samples were diluted 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. Real-time PCR reactions were performed as described previously [5]. The primers used in this experiment (*rbcL-a, matK, rpoB, rpoC1, trnL-F, trnH-psbA*) have been described previously [2].

Equal amounts of genomic DNA from three species were used to create the mixed-species metabarcoding templates. Sequencing reactions comprised nine species.

2.3 qPCR efficiency and Cq calculation

qPCR efficiency and Cq were computed using qpcR, R package. Efficiency value (E) was calculated as EcpD2=F(cpD2)/F(cpD2)-1 [6].

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

PCR products for *rbcL-a*, *rpoC1* and *rpoB* primers from mixed genomic DNAs were pooled equivalently to yield a final amount of 100ng. Preparation of samples for library construction and sequencing were performed using the Ion Torrent Next generation sequencing Kits (Life Technologies, USA). 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 330bp. Emulsion PCR was performed using One-touch system and sequencing was performed using 314 Ion Torrent chips. A total of 333,274 reads 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 [7]. 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].

3. Results and discussion

3.1. Suitability of barcodes and qPCR parameters for specific barcodes depending on plant species

Our first analysis assessed PCR success. 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 low PCR success, 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 [3]. 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.

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 q-PCR varied between barcode markers. The highest average efficiency, based on amplification from all species, corresponded to the markers *trnL–F* and *trnH-psbA*. 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 different from high-efficiency marker *trnL-F* ($p<0.0001$ for *matK* and *rbcL* and $p=0.0013$ for *rpoC1*).

Looking at intra-species variation for all barcodes, Cq values varied widely also in this case (Fig. 2). 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 [9], and assumed an average efficiency for both markers (efficiency = 1.9), the predicted differences in starting DNA level would be 2116-fold based on the estimates from these two barcodes.

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.

Our results show that differences in PCR 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.

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.

3.3. 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 amplify in each species. The results of simultaneous sequencing of equal amounts of PCR products from mixed species templates amplified with barcode markers, *rbcL*, *rpoB* and *rpoC1* 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 *Solanum tuberosum* and only 0.02% to *Zea 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.

Analysis of read numbers also showed a strong bias in the number of total reads corresponding to each of the barcodes. 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 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*.

4. Conclusions

Our results show that bias existing in PCR amplification and NGS can interfere with correct, especially quantitative, analysis of matabarcoding samples. 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.

5. Acknowledgments

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”. This work was published in Analytical and Bioanalytical Chemistry (407(7): 1841-8)

6. References

[1] Hajibabaei M., Singer G., Hebert P., Hickey D. 2007 DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. Trends Genet 23:167–172.

[2] Hollingsworth P., Forrest L., Spouge J., et al. 2009 A DNA barcode for land plants. Proc Natl Acad Sci U S A 106:12794–12797

[3] Kress W., Erickson D. 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.

[4] Schmittgen T., Livak K. 2008 Analyzing real time PCR data by the comparative CT method. Nat Protoc 3:1101–1108.

[5] Mallona I., Weiss J., Egea-Cortines M. 2011 pcrEfficiency: a Web tool for PCR amplification efficiency prediction. BMC Bioinformatics 12:404.

[6] Spiess A.-N., Feig C., Ritz C. 2008 Highly accurate sigmoidal fitting of real-time PCR data by introducing a parameter for asymmetry. BMC Bioinformatics 9:221.

[7] Polz M., Cavanaugh C. 1998 Bias in Template-to-Product Ratios in Multitemplate PCR. Appl Envir Microbiol 64:3724–3730.

[8] Links M., Chaban B., Hemmingsen S., 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

[9] Schmittgen T., Livak K. 2008 Analyzing real-time PCR data by the comparative CT method. Nat Protoc 3:1101–1108

Tables and Figures

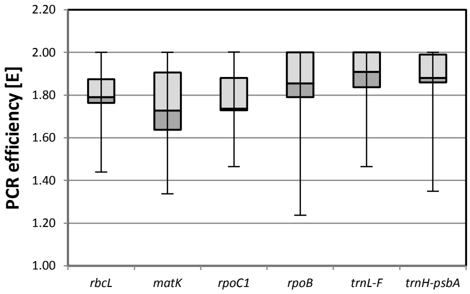

Figure 1. *Boxplot of PCR efficiency data for six barcoding markers derived from qPCRs of 48 plant species. The graphic shows only successful amplification data with an efficiency >1*

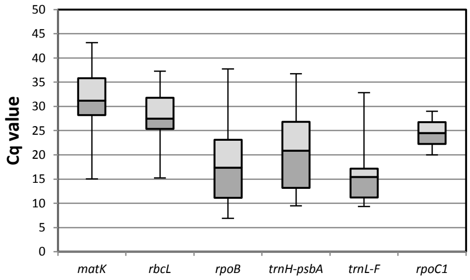

Figure 2. *Boxplot of Cq values for six barcoding markers derived from qPCRs of 48 plant species*