Quantitative evaluation of bias in barcode markers derived from complex samples

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

[(1)] Dirección Genetics, Instituto de Biotecnología Vegetal, Universidad Politécnica de Cartagena, 30202, Cartagena, España. 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, España

ABSTRACT

PCR products have become a major commodity used to identify organisms based on polymorphism at the DNA level. One problem arising is that unbiased identification of organisms takes as working hypothesis that when DNA is extracted from a sample, a positive signal will be obtained if universal primers are used and DNA quality is suitable for PCR. As this assumption is not always correct we used a system where large differences in PCR success have been described to identify where biases appear and maybe identify solutions. Plants can be identified with at least seven independent plastid-located loci. These differ in their degree of PCR success and how informative they are in terms of taxonomically useful sequence polymorphisms. Here we used six common plastid loci spanning 48 plant species and performed a quantitative analysis of bias at each step of the identification process. As expected we found important differences in PCR efficiency within a single species, depending on the barcoding sequence being amplified. Quantitative PCR revealed that the Ct threshold for various plastid loci, even within a single species, could exhibit greater than 2000-fold differences in DNA quantity after amplification. We then performed Next Generation Sequencing experiments in nine species using equal quantities of three plastid-based primers and equally-mixed quantities of DNA from multiple species. The result was significantly biased towards species and specific loci even when using adaptor-specific primers. Our results caution that Next-Generation Sequencing projects may suffer dramatic bias, arising largely during DNA amplification steps. Moreover, that amplification-based Next Generation Sequencing technologies exhibit additional bias despite using adaptor-specific primers, indicating that amplification success depends on the DNA fragment. As such, while qualitative analysis of unknown samples are prone to false negative results if a combination of widely-successful amplicons are not used, quantitative results should be considered highly suspect, even if all species in the starting sample are known.

Keywords: meta-barcoding, Next Generation Sequencing, Ion torrent, PCR efficiency

1. Introduction

DNA barcoding is an important tool in taxonomic discrimination basing on universal DNA sequences [1]. These are using in monitoring species distribution in biodiversity studies and biodiversity conservation. The barcoding approach has become a mainstream technique to identify species in insects [2], very closely related plant species or hybrids [3] and bacteria [4].

In plants it is not easy to find a single locus that can be used as an efficient barcode as it is in animals or bacteria. Until now the most promising candidate markers are *matK* [5] and the combination of *trnH-psbA* and *rbcL*1 [6].

Identification of environmental samples comprising more than one species is very difficult, especially when this kind of analysis aim to arrive at a quantitative measure of the relative abundance of the various species in the sample.

There are different types of biases that can influence results of identification. Some of them as DNA extraction efficiency, chloroplast copy number [7], etc. However some other sources of bias can be used to improve experimental

approach for a given sample, and/or better normalize the outcomes.

These biases fall into three categories: differential barcode amplification success as a result of the barcode's universal primers, 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 and biases introduced during the preparation of DNA libraries for sequencing.

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 data to better-reflect the population make-up.

The present study, therefore, aims to first quantitatively analyze PCR success and evaluate amplification efficiency and Ct values as a tool for predicting amplification success of specific barcoding markers. 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 simultanously 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.

2. Materials y Methods

2.1 Plant material, DNA extraction and real-time PCRPlant material was gathered from the local fruit market, field sampling, botanical records and our own collections. Fresh leaf material from 48 plant species belonging to 33 different families was used for further analysis. Two independent total genomic DNA samples were extracted from fresh leaf using the commercial kit 'Plant NucleoSpin' (Machery and Nagel, Düren, Germany). 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/reaction. Real-time PCR reactions were performed with the Mx3000P QPCR System using the SYBR Premix ExTaqTM (Takara Biotechnology, Dalian, China) with ROX as a reference dye.

Equal amounts of genomic DNA from three species were used to create the mixed-species templates. Amplifications were performed using the same protocol described above, except that the initial DNA quantity was 150 ng corresponding to 50ng of each of the three genomes. Sequencing reactions comprised nine species.

2.3. qPCR efficiency

qPCR efficiency was computed using qpcR, R package. Efficiency value (E) was calculated as EcpD2=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 [8].

2.4. 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 100ng. 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. A total of 333,274 reads with a mean read length of 159bp were computationally analyzed in order to identify species origin of each fragment. Such analysis was performed by aligning the reads with a library of known Chloroplast sequences using Bowtie2 [9] , and then extracting from the resulting SAM file a map of reads to the known chloroplast sequences, using a Perl script from the mPuma pipeline [10].

3. Results and Discussion

Our first analysis was geared towards an assessment of 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 (Fig.1). Thus, although *matK* is highly informative in its ability to discriminate between related taxa, it could cause major biases.

The second phase of the analysis addressed whether end point PCR results are the outcome of PCR efficiency. As shown in Fig. 2, 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 different from high-efficiency marker *trnL-F* ($p<0.0001$ for *matK* and *rbc*L and $p=0.0013$ for *rpoC1*).

Differences in efficiency 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.

The identification of genomic DNAs corresponding to different organisms in environmental samples requires sequencing of barcode-PCR products. The result 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 to 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.*

Analysis of read numbers also showed a strong bias in the number of total reads corresponding to each of the marker. 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 $p< 2.2 e\text{-}16$).

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 universal markers, quantitated our results using widely-accepted practices such as qPCR, and followed normal protocols for library construction and Next-generation sequencing. At each stage, we re-normalized the samples such that we knew the precise quantities and relative abundances of the input DNA.

Similarity between primer and template [11], as well as the regional G+C content of a template, are factors that influence PCR efficiency [12]. 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 marker exist [6].

Late development towards short barcodes [13] certainly improve the current situation, but false negatives still remain an issue. Furthermore, even an experimental design geared towards qualitative identification of an environmental sample of unknown composition could exhibit numerous false negative results unless multiple barcode markers are used.

4. Conclusions

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 sequence analysis from environmental DNA samples. Careful consideration of pre-amplification and sequencing methods might minimize these biases, though we suggest that they cannot be completely overcome using current barcoding technologies relying strongly on PCR amplification.

5. Acknowledgments

This work was funded by the Comunidad Autónoma de la Región de Murcia Project “Molecular markers in coservation and management of the flora of Murcia Region”.

6. References

[1] Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. Trends in Genetics, 23, 167–172

[2] Burns JM, Janzen DH, Hajibabaei M, Hallwachs W, Hebert PDN (2008) DNA barcodes and cryptic species of skipper butterflies in the genus Perichares in Area de Conservacion Guanacaste, Costa Rica. Proceedings of the National Academy of Sciences of the United States of America, 105, 6350–6355

[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 Systematics and Evolution, 298, 409–419

[4] 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

[5] Hollingsworth PM, Forrest LL, Spouge JL et al. (2009) A DNA barcode for land plants. Proceedings of the National Academy of Sciences of the United States of America, 106, 12794–12797

[6] 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

[7] Rowan BA, Bendich AJ (2009) The loss of DNA from chloroplasts as leaves mature: fact or artefact? Journal of experimental botany, 60, 3005–10

[8] 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

[9] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature methods, 9, 357–9

[10] Links MG, Chaban B, Hemmingsen SM, Muirhead K, Hill JE (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

[11] Mann T, Humbert R, Dorschner M, Stamatoyannopoulos J, Noble WS (2009) A thermodynamic approach to PCR primer design. Nucleic acids research, 37, e95

[12] Suzuki M, Giovannoni S (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl. Envir. Microbiol., 62, 625–630

[13] Taberlet P, Coissac E, Pompanon F et al. (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. Nucleic Acids Res, 35, e14

Tables and Figures

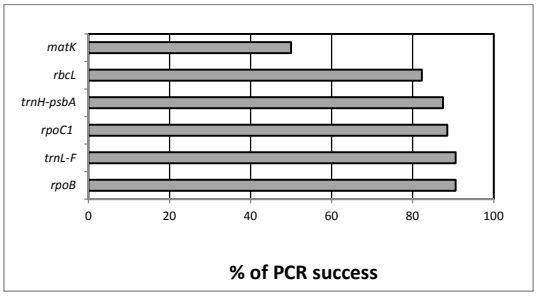

Figure 1. *Percent PCR success of six barcoding markers in a survey of 48 plant species*

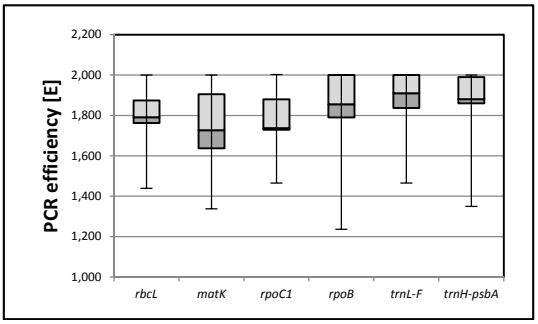

Figure 2. *Boxplot of PCR efficiency data for six barcoding markers derived from qPCRs of 48 plant species. The graphic shows only successful amplification data with an efficiency > 1*