# Comment on Bohmann et al. Strategies for sample labelling and library preparation in DNA metabarcoding studies

Peter Hamback[1], Jasmina Sargac[2], and Magdalena Grudzinska-Sterno[1]

[1]Stockholm University
[2]Swedish University of Agricultural Sciences Faculty of Natural Resources and Agricultural Sciences

March 31, 2022

**Abstract**

DNA metabarcoding necessitates labelling amplicons in order to connect sequencing reads with samples, but labelling protocols may cause errors where indexes are incorrectly assembled during PCR due to tag-jumping. A recent paper by Bohmann et al (2021) reviews the main labelling methods and point out that library building using PCR's on tagged amplicons may be particularly problematic. Due to unforeseen problems in two sequencing projects, we had to use a second PCR on tagged amplicons to salvage two large data sets. This test showed that the problems with tag-jumping errors were acceptable and could be accounted for during analysis, if handled properly when designing the indexing strategy.

Comment on Bohmann et al. Strategies for sample labelling and library preparation in DNA metabarcoding studies

Hambäck, P.A.[1*], J. Sargac[2] and M. Grudzinska-Sterno[1]

[1]Department of Ecology, Environment and Plant Sciences, Stockholm University; [2]Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences

Corresponding author: Department of Ecology, Environment and Plant Sciences, Stokholm University, 106 91 Stockholm, Sweden. E-mail: peter.hamback@su.se

Abstract

DNA metabarcoding necessitates labelling amplicons in order to connect sequencing reads with samples, but labelling protocols may cause errors where indexes are incorrectly assembled during PCR due to tag-jumping. A recent paper by Bohmann et al (2021) reviews the main labelling methods and point out that library building using PCR's on tagged amplicons may be particularly problematic. Due to unforeseen problems in two sequencing projects, we had to use a second PCR on tagged amplicons to salvage two large data sets. This test showed that the problems with tag-jumping errors were acceptable and could be accounted for during analysis, if handled properly when designing the indexing strategy.

## Introduction

DNA metabarcoding has rapidly become a mainstream tool in ecological research, both for species inventories using environmental DNA and for identification of gut contents when quantifying diets (Liu, Clarke, Baker, Jordan, & Burridge, 2020), not only because of the reduced costs for sequencing but also because of the development of protocols that increase the quality of output data (reviewed in Alberdi et al., 2019; Alberdi, Aizpurua, Gilbert, & Bohmann, 2018). Because metabarcoding comes with contamination risks during both data collection and laboratory procedures, several papers describe methods to reduce contamination during

field collections and preparation of samples before sequencing (Alberdi et al., 2018; King et al., 2012; King, Read, Traugott, & Symondson, 2008). One recent issue under discussion concerns labelling of amplicons to enable the connection of samples and sequence data (Schnell, Bohmann, & Gilbert, 2015). To enable the identification of metabarcoding data, the workflow involves methods of labelling amplicons through the addition of nucleotide tags on the 5'-end of metabarcoding primers and/or as indexes during library preparation. There are three main strategies for labelling in metabarcoding studies, of which one (the so-called 'tagged PCR approach') can result in tag-jumps, i.e., the appearance of sequences carrying new combinations of the used 5' nucleotide tags (e.g. Esling et al 2015, Schnell et al 2015). A recent review by Bohmann et al. (2021) nicely describe the risks connected to different indexing methods, each with their different pros and cons.

As reviewed by Bohmann et al. (2021), risks of getting erroneous indexed tag combinations will only occur as a result of library preparation with T4 DNA Polymerase blunt-ending or when libraries are prepared by ligating indexes on pools of tagged amplicons in a second PCR, suggesting that these methods should be avoided (Carøe & Bohmann, 2020). While we completely agree with this general advice, problems during execution may necessitate workflow changes involving a second PCR to salvage data. Moreover, including a risk analysis before setting up the workflow may allow for a smooth transition between methods if needed. In our group, we use metabarcoding to identify prey in spider guts using the tagged PCR approach with metabarcoding primers having 5' nucleotide tags (Binladen et al., 2007) (hereafter tags). We thereafter pool amplicons and build libraries using a PCR-free protocol that includes a ligation of Illumina adaptors with dual indexes (for complete laboratory protocols see Hambäck et al., 2021). As Bohmann et al. (2021) points out, this approach cannot cause tag jumping if a blunt-ending step is excluded from the library preparation protocol. For this reason, we omit the end-repair step and perform phosphorylation and adenylation of DNA fragments in a separate reaction.

A problem when extracting DNA from guts of small organisms is that prey DNA is in low amounts and highly fragmented. In two recent projects, even after PCR amplification, our samples were found by the sequencing lab to contain too little DNA for the MiSeq to run properly. At this stage, we either had to abandon the projects, wasting resources in collecting and preparing the samples, or use additional PCR steps on tagged amplicon pools to boost DNA amounts. Following discussions with the staff at our National Genomics Institute (NGI), our sequencing facility, we decided to run additional PCRs with only 6 cycles. Libraries were prepared using SMARTer ThruPLEX DNA-seq library preparation kit excluding fragmentation of DNA (Takara Bio), and to measure tag jumping rates we only used 75% of available tag combinations while leaving about 25% empty. Moreover, to force tag jumping errors to occur only between spider individuals within sites, the new libraries were reconstructed based on sampling units. Because downstream analyses in this study focused on between site differences and therefore data were pooled over spider individuals, tag jumping between spiders within sites did not distort the result. This unplanned library adjustment necessitated some spider individuals to be discarded to avoid duplicity of tags between individuals within sites. In retrospect, this problem could have been avoided when setting tags to samples.

In post-processing, we used standard settings to clean, filter, de-multiplex and tabulate sequence data using Obitools (Boyer et al., 2016) on the Galaxy platform (Jalili et al., 2020) similar to our previous studies. Thus, pair-end reads were joined using 'Illuminapairedend', trimmed and annotated using 'NGSfilter' before filtering on length using 'obigrep' (310-330 bp) and identifying unique sequences with 'obiuniq'. After tabulating the data, we identified OTU's using 'pick_otus' and connected representative sequences based on 'pick_rep_set" to taxon identities using Barcode of Life Database (BoLD) (Ratnasingham & Hebert, 2007). Here we only report sequence number distributions separated between correct and false tag combinations, leaving results on actual spider diets to other publications.

In project 1, involving linyphiid spiders, the yield of useful prey sequences was about 8.7 million sequences after cleaning and demultiplexing. Among these sequences, only 0.03% were connected to false tag combinations. Because about 25% of tag combinations were empty, the tag jumping rate can be estimated to be about 0.12%. In project 2, involving lycosid spiders, the yield was lower (0.45 million sequences), with a

higher percentage (1.9%) of sequences connected to false tag combinations, corresponding to an approximate tag jumping rate of 7.6%. Notable is that the tag jumping rate decreased to about 5.8% when excluding sequences with no match in BoLD. When examining the frequency distribution of sequence number for true and false tag combinations (Fig. 1), it was also apparent that there was almost no overlap in the distributions for project 1 but a larger overlap in project 2.

For the data from project 2, we decided to further compare proportions of false combinations between the 32 libraries. It was evident that these proportions showed large variation between libraries, from 0.004% up to 30%. The reason for this variability is not evident, but data suggest that problems with high frequencies of false combinations mainly occurred at a low total yield per library, in our case below 5000 sequences (Fig. 2), which could explain why no problems appeared in project 1. The reason for the low yield in project 2 is unclear, but it served us well to illustrate the yield dependent error rates. It is apparent that for this protocol, tag jumping errors are unproblematic as long as yield is sufficiently high. For the final diet analyses, we will use estimated tag jumping rates to set dynamic thresholds for data exclusions at the level of sampling sites and species (see Cirtwill & Hambäck, 2021).

To summarize, similar to previous studies (Carøe & Bohmann, 2020; Schnell et al., 2015), we find that tag jumping errors are potential problems in metabarcoding studies when libraries are built using a PCR on pools of tagged amplicons and such an approach should be avoided when possible. However, as in our case, even when using a PCR-free library preparation protocol it is sometimes necessary to enrich libraries to obtain sufficient concentrations for sequencing. In an ideal world, we could have collected new samples but costs are often prohibitive. We instead used a strategy where error rates could be estimated and where effects from errors could be avoided. When doing this, we find that estimated error rates due to tag jumping are small when DNA yields are not very low, suggesting that risks of enriching tagged amplicon libraries through additional PCR cycles can be acceptable. This information is good news when aiming to describe gut contents of small invertebrate predators, where sometimes DNA amounts can be very low. However, it is advisable to consider risks prior to designing tagging protocols to enable future methodological switches.

Acknowledgements

Y. Marincevic-Zuniga and R. Kudva was very helpful in trouble-shooting, whereas K. Bohmann and R.K. Johnson provided helpful comments on a previous version of this manuscript. Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala. This facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported but the Swedish Research Council and the Knut and Alice Wallenberg Foundation.

References

Alberdi, A., Aizpurua, O., Bohmann, K., Gopalakrishnan, S., Lynggaard, C., Nielsen, M., & Gilbert, M. T. P. (2019). Promises and pitfalls of using high-throughput sequencing for diet analysis. *Molecular Ecology Resources, 19* (2), 327-348. doi:10.1111/1755-0998.12960

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution, 9* (1), 134-147. doi:10.1111/2041-210x.12849

Binladen, J., Gilbert, M. T. P., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R., & Willerslev, E. (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One, 2* , e197. doi:10.1371/journal.pone.0000197

Bohmann, K., Elbrecht, V., Carøe, C., Bista, I., Leese, F., Bunce, M., . . . Creer, S. (2021). Strategies for sample labelling and library preparation in DNA metabarcoding studies. *Molecular Ecology Resources* . doi:10.1111/1755-0998.13512

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. *Molecular Ecology Resources, 16* , 176-182.

doi:10.1111/1755-0998.12428

Carøe, C., & Bohmann, K. (2020). Tagsteady: A metabarcoding library preparation protocol to avoid false assignment of sequences to samples.*Molecular Ecology Resources, 20* (6), 1620-1631. doi:10.1111/1755-0998.13227

Cirtwill, A. R., & Hambäck, P. (2021). Building food networks from molecular data: Bayesian or fixed-number thresholds for including links.*Basic and Applied Ecology, 50* , 67-76. doi:10.1016/j.baae.2020.11.007

Hambäck, P. A., Cirtwill, A. R., García, D., Grudzinska-Sterno, M., Miñarro, M., Tasin, M., . . . Samnegård, U. (2021). More intraguild prey than pest species in arachnid diets may compromise biological control in apple orchards. *Basic and Applied Ecology, 57* , 1-13. doi:10.1016/j.baae.2021.09.006

Jalili, V., Afgan, E., Gu, Q., Clements, D., Blankenberg, D., Goecks, J., . . . Nekrutenko, A. (2020). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update.*Nucleic Acids Res, 48* (W1), W395-W402. doi:10.1093/nar/gkaa434

King, R. A., Davey, J. S., Bell, J. R., Read, D. S., Bohan, D. A., & Symondson, W. O. C. (2012). Suction sampling as a significant source of error in molecular analysis of predator diets. *Bulletin of Entomological Research, 102* (3), 261-266. doi:10.1017/S0007485311000575

King, R. A., Read, D. S., Traugott, M., & Symondson, W. O. C. (2008). Molecular analysis of predation: a review of best practice for DNA-based approaches. *Molecular Ecology, 17* (4), 947-963. doi:10.1111/j.1365-294X.2007.03613.x

Liu, M. X., Clarke, L. J., Baker, S. C., Jordan, G. J., & Burridge, C. P. (2020). A practical guide to DNA metabarcoding for entomological ecologists. *Ecological Entomology, 45* , 373-385. doi:10.1111/een.12831

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (*www.barcodinglife.org*).*Molecular Ecology Notes, 7* (3), 355-364. doi:10.1111/j.1471-8286.2007.01678.x

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources, 15* (6), 1289-1303. doi:10.1111/1755-0998.12402

Figure legends

Fig. 1. Frequency distribution of sequence number among spider individuals with correct (filled bars) and false (open bars) tags in two separate data sets (A, B). Notice that false tags have a similar distribution for the two data sets but that correct tags are much lower in (B).

Fig. 2. Relationship between the proportion of sequences with false tag combinations and total DNA yield per library (N=32).

Figure 1

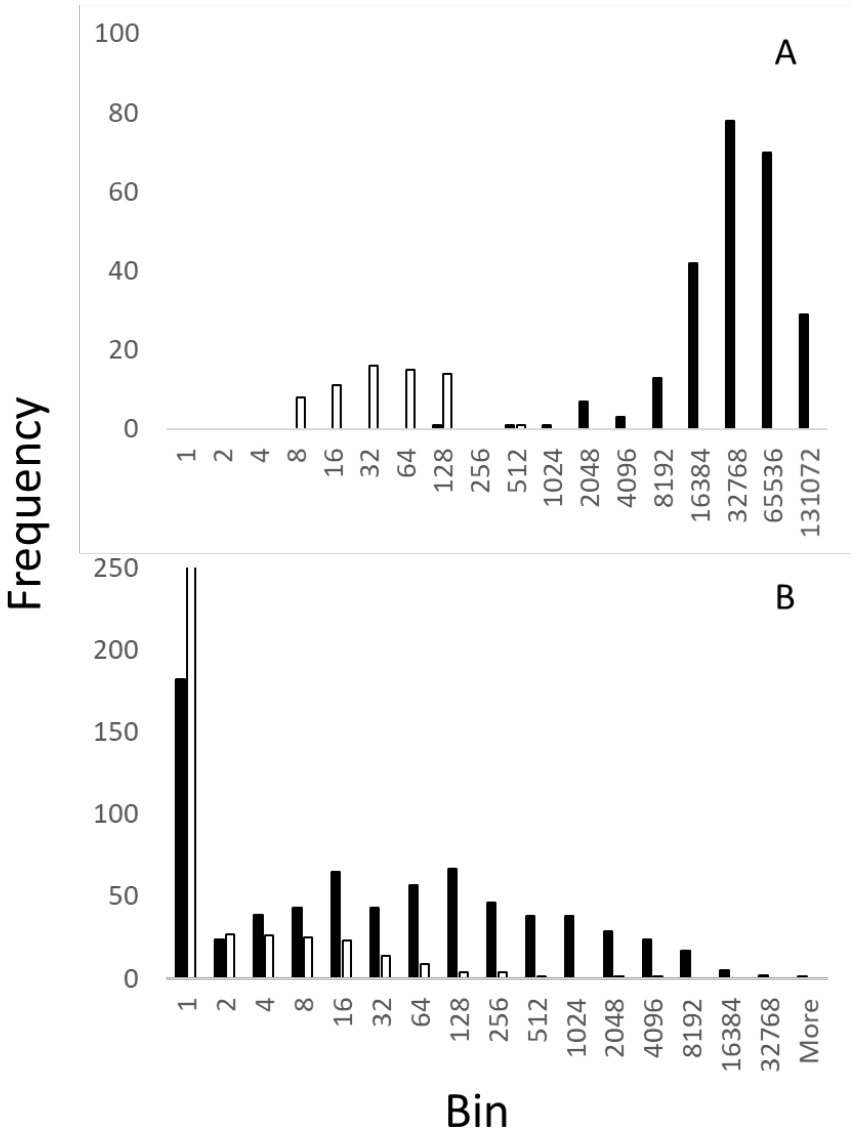


Figure 2