FUNGAL ECOLOGY XXX (2012) 1–3

available at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/funeco

Commentary

Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies?

SUMMARY

High throughput sequencing has become a powerful tool for fungal ecologists to explore the diversity and composition of fungal communities. However, various biases and errors are associated with the new sequencing techniques that must be handled properly. We here provide evidence for a source of error that has not yet been taken into account.

During amplicon pyrosequencing we incorporate tags in both ends of the amplicons, which allows us to check for tag coherence after sequencing. In several studies we have observed that a small proportion of the resulting sequences possess novel tag combinations. Our observations cannot be explained by primer contamination or PCR chimaeras. This indicates that some DNA fragments switch tags during laboratory setup. If not controlled for, this will cause numerous false positives in downstream analyses. In most amplicon pyrosequencing studies of fungal communities, amplicons are typically tagged in one end only. We suggest that amplicons should be tagged in both ends before pyrosequencing to control for tag switching.

High throughput sequencing (HTS) approaches, such as 454 pyrosequencing (Margulies *et al.* 2005), have provided new and exciting opportunities to explore the community composition and diversity of micro- (e.g. Buée *et al.* 2009; Jumpponen & Jones 2009; Blaalid *et al.* 2012) and macro-organisms (e.g. Epp *et al.* 2012) in environmental samples. In amplicon HTS studies, a specific DNA region from one or a few groups of organisms are PCR amplified by group selective primers prior to 454 pyrosequencing.

Various types of errors may be incorporated during PCR and DNA sequencing that need to be accounted for to avoid inflated species richness estimates (Kunin *et al.* 2010) or the inclusion of artificial lineages (Berney *et al.* 2004). Even high fidelity polymerases will have nucleotide incorporation errors during the PCR step producing artificial mutations in some of the amplified fragments (Qiu *et al.* 2001; Li *et al.* 2006). Moreover, chimeric DNA fragments that combine different parts of different DNA templates are often generated during PCR (Qiu *et al.* 2001; Lahr & Katz 2009). Additional errors are added during sequencing. For example, the 454 sequencing technique fails to handle long homopolymer stretches satisfactorily, frequently leading to an incorrect number of incorporated bases in such regions.

In HTS studies, numerous samples are normally tagged, pooled and sequenced in parallel on the same illumina/454 plate or section of a plate (Binladen *et al.* 2007; Hamady *et al.* 2008). The unique tags (also named molecular identifiers (MIDs)) are used to link the output reads to the original samples after sequencing. The tags usually consist of 6–10 base pairs. These are either linked to the amplified fragments directly during PCR or ligated after PCR. In most published amplicon pyrosequencing studies of fungi, the amplified fragments are only tagged on one end (e.g. Buée *et al.* 2009; Jumpponen & Jones 2009).

In our amplicon pyrosequencing studies we routinely tag the amplicons on both ends. This enables us to control for the presence of non-compatible tag combinations after sequencing using the bioinformatics pipeline CLOTU (Kumar *et al.* 2011). In several independent amplicon pyrosequencing studies we have observed and reported output reads with non-compatible tag combinations (Aas 2010; Blaalid *et al.* 2012; Kauserud *et al.* 2012; plus several ongoing unpublished studies). On the basis of two of these datasets (Aas 2010 and a yet unpublished work led by Daniel Lindner), we suggest that tag switching might be a common but largely overlooked phenomenon in amplicon pyrosequencing studies. Clearly, this type of error should be considered, and when possible, controlled for.

In one study (Lindner *et al.* in prep), a total of 176 247 ITS1 sequences were generated from 99 axenic single spore cultures representing a similar number of different fungal species, where each taxon (= sample) was labeled with a unique tag. The experimental setup for 454 sequencing followed Blaalid *et al.* (2012). Tags were added to both ends of the fragment by using fusion primers containing the adaptor sequence for emulsion PCR, a MID and the PCR primer. We used a subset of the recommended Roche multiplex identifiers (technical bulletin 005–2009) where the MID sequence is 10 bp long and at least four changes (insertion, deletion, substitution) different from the other members of the MID set. Samples were standardized using SequalPrep™ Normalization Plate (96) Kit following the manufacturer's protocol (Invitrogen, CA, USA) and cleaned with Wizard® SV Gel and PCR Clean-Up System (Promega, Madison, USA). Samples were run on a bioanalyzer to check for the successful removal of all primers and primer dimers before emPCR.

Please cite this article in press as: Carlsen T, et al., Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies?, Fungal Ecology (2012), http://dx.doi.org/10.1016/j.funeco.2012.06.003

When only the forward tag was used for assigning sequences to samples (taxa), we observed that many samples included a low abundance of 'contaminant sequences' from one or several of the other co-analyzed samples (taxa). These "contaminant" sequences represented taxa mixed into the same tube prior to emulsion PCR (emPCR) and run on the same 454 lane. When searching these contaminant sequences for the reverse tag, we observed that they had non-compatible tag combinations where the reverse tag indeed was the expected "correct" tag for the sample in question. As a consequence of this observation, we also controlled the reverse tag of sequences with the correct forward tag. We found that 0.7 % of the sequences with a correct forward tag had a non-compatible reverse tag from another sample in the same sequencing lane. In total, about 1.6 % of the full-length ITS1 reads in this dataset had a non-compatible tag combination. Noteworthy, these reads did not represent PCR chimeras as only the tags had been switched. Lab contamination is also not likely as there were samples PCR amplified separately, from different mastermixes, but pooled to be run on the same 454 compartment that had switched tags. Samples run side by side on a PCR plate but run on different compartments on the 454 plate showed no cross contamination and no tag switch.

In another dataset (Aas 2010), the diversity of phylloplane fungi associated with the grass species *Avenella flexuosa* was analyzed using amplicon pyrosequencing of the ITS1 region. Ninety plant samples were analyzed using the same setup for 454 sequencing as above. Here we observed that 2.3 % of the raw reads exhibited non-compatible tag combinations.

We want to emphasize that the observed amplicons with non-matching tags cannot be explained by PCR chimaeras as these samples were not PCR amplified together. Furthermore, a mixed template on an emulsion PCR bead would give a nonsense read completely different from our observed results. Sequencing error is also not plausible as we have used MID that are at least four changes different from the other members of the MID set, and we have not allowed for any MID errors during filtering.

The PCR amplicons could theoretically have switched tags either prior to emulsion PCR when mixed into the same tube, or during the emPCR. We presume that the first scenario is most likely. Even after thorough rinsing of PCR amplicons with kits such as Agencourt® AMPure® (Beckman Coulter, High Wycombe, UK) or Wizard® SV Gel and PCR Clean-Up System (Promega, Madison, USA), that we used, low concentrations of unused tagged primers may be available in the solution and may interfere with the ITS1 fragments prior to emPCR. Another possibility is that the primer/adaptor regions of amplicons from different samples align and cross during a denaturation step. We would encourage researchers to look for tag switching in their datasets so that the research community eventually will be able to find the cause of this effect.

Our results demonstrate that reliance on only one tag during amplicon HTS may lead to undetected mixing of sequences across samples and consequently numerous false positives. In accord with our findings, van Orsouw *et al.* (2007) observed from 0.1 to 16 % fragments with non-compatible tag combinations in a 454 sequencing setup with mixed samples. Although the number of reads with non-compatible tag combinations were low in the referred studies, it may severely impact the downstream analyses if not properly corrected for. For example, tag switching may create an artificial evenness of samples pooled in a single lane or plate compartment, especially if the data are treated as presence/absence. Precautionary measures that can be taken to avoid tag switching include thorough rinsing of PCR products, cold storage of pooled amplicon libraries immediately after mixing, and reduced sample storage time between the final steps in the laboratory preparations. Removing low-frequency OTUs would counteract the bias to some extent, but only if performed per individual sample. This is because a mis-tagged fragment in one sample is likely abundant in another sample. We also advocate randomizing samples to avoid all samples from, for example, one treatment or plot being pooled in one lane. We think this may be a general problem in all HTS amplicon studies where samples are tagged before pooling as it was also apparent in van Orsouw *et al.* (2007), where the tags were added by ligation.

Unfortunately, the commonly used software programs for processing high throughput amplicon sequence data, such as Qiime (Caporaso *et al.* 2010) and Mothur (Schloss *et al.* 2009) do not include satisfactory options for the control of non-compatible tag combinations. However, such options may be easily implemented, as done for instance in the software program CLOTU (Kumar *et al.* 2011). Cross contamination and sample mix ups in HTS studies might be a common but largely overlooked phenomenon that has possibly had major impact in some studies, as recently demonstrated by Westra *et al.* (2011).

REFERENCES

Aas AB, 2010. *Diversity and Species Composition of Fungal Endophytes in* Avenella flexuosa *under Different Sheep Grazing Regiments*. Master thesis in Biology, University of Oslo.

Berney C, Fahrni J, Pawlowski J, 2004. How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys. *BMC Biology* **2**: 13.

Binladen J, Gilbert MT, Bollback JP, 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.

Blaalid R, Carlsen T, Kumar S, Halvorsen R, Ugland KI, Fontana G, Kauserud H, 2012. Changes in the root associated fungal communities along a primary succession gradient analyzed by 454 pyrosequencing. *Molecular Ecology* **21**: 1897–1908.

Buée M, Reich M, Murat C, Morin E, Nilsson RH, Uroz S, Martin F, 2009. 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist* **184**: 449–456.

Caporaso JG, Kuczynski J, Stombaugh J, et al., 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**: 335–336.

Epp LS, Boessenkool S, Bellemain EP, Haile J, Esposito A, Riaz T, ErsÉUs C, Gusarov VI, Edwards ME, Johnsen A, Stenøien H, Hassel K, Kauserud H, Yoccoz N, Bråthen KA, Willerslev E, Taberlet P, Coissac E, Brochmann C, 2012. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. *Molecular Ecology* **21**: 1821–1833.

Please cite this article in press as: Carlsen T, et al., Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies?, Fungal Ecology (2012), http://dx.doi.org/10.1016/j.funeco.2012.06.003

Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R, 2008. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods* **5**: 235–237.

Jumpponen A, Jones KL, 2009. Massively parallel 454 sequencing indicates hyperdiverse fungal communities in temperate *Quercus macrocarpa* phyllosphere. *New Phytologist* **184**: 438–448.

Kauserud H, Kumar S, Brysting A, Nordén J, Carlsen T, 2012. High consistency between replicate 454 pyrosequencing analyses of ectomycorrhizal plant root samples. *Mycorrhiza* **22**: 309–315.

Kumar S, Carlsen T, Mevik B-H, Enger P, Blaalid R, Shalchian-Tabrizi K, Kauserud H, 2011. CLOTU: an online pipeline for processing and clustering of 454 amplicon reads into OTUs followed by taxonomic annotation. *BMC Bioinformatics* **12**: 182.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P, 2010. Wrinkles in the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity estimates. *Environmental Microbiology* **12**: 118–123.

Lahr DJ, Katz LA, 2009. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques* **47**: 857–866.

Li M, Diehl F, Dressman D, Vogelstein B, Kinzler KW, 2006. BEAMing up for detection and quantification of rare sequence variants. *Nature Methods* **3**: 95–97.

Lindner D, Carlsen T, Nilsson H, Schumacher T, Kauserud H. 454 amplicon sequencing reveals intragenomic ITS divergence in fungi., in prep.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–380.

Qiu X, Wu L, Huang H, McDonel PE, Palumbo AV, Tiedje JM, Zhou J, 2001. Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16 rRNA gene-based cloning. *Applied and Environmental Microbiology* **67**: 880–887.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF, 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**: 7537–7541.

van Orsouw NJ, Hogers RCJ, Janssen A, Yalcin F, Snoeijers S, Verstege E, Schneiders H, Hvd Poel, Jv Oeveren, Verstege H, Schneiders H, van der Poel H, van Oeveren J, Verstegen H, van Eijk MJT, 2007. Complexity reduction of polymorphic sequences (CRoPS™): A novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE* **2**: e1172.

Westra H-J, Jansen RC, Fehrmann RSN, te Meerman GJ, van Heel D, Wijmenga C, Franke L, 2011. MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* **27**: 2104–2111.

ARTICLE INFO

Accepted 8 June 2012

Corresponding editor: Petr Baldrian

KEYWORDS

Amplicon pyrosequencing
Diversity
Fungi
High throughput sequencing
Sequencing errors

***Tor CARLSEN*[a,*], *Anders Bjørnsgaard AAS*[a], *Daniel LINDNER*[b], *Trude VRÅLSTAD*[a], *Trond SCHUMACHER*[a], *Håvard KAUSERUD*[a]**

[a]*Microbial Evolution Research Group (MERG), Department of Biology, University of Oslo, P.O. Box 1066 Blindern, NO-0316 Oslo, Norway*
[b]*US Forest Service, Northern Research Station, Center for Forest Mycology Research, One Gifford Pinchot Drive, Madison, WI 53726, USA*
**Corresponding author*. Tel.: +47 22854588.
E-mail address: tor.carlsen@bio.uio.no(T. Carlsen).

1754-5048/$ – see front matter

http://dx.doi.org/10.1016/j.funeco.2012.06.003