Evaluation of a new primer combination to minimize plastid contamination in 16S rDNA metabarcoding analyses of alga-associated bacterial communities

François Thomas, Simon Dittami, Maéva Brunet, Nolwen Le Duff, Gwenn Tanguy, Catherine Leblanc, Angélique Gobet

environmental microbiology
environmental microbiology reports

Evaluation of a new primer combination to minimize plastid contamination in 16S rDNA metabarcoding analyses of alga-associated bacterial communities

Journal:	*Environmental Microbiology and Environmental Microbiology Reports*
Manuscript ID	EMIR-2019-1373.R1
Journal:	Environmental Microbiology Reports
Manuscript Type:	EMIR - Brief report
Date Submitted by the Author:	n/a
Complete List of Authors:	Thomas, François; Station Biologique de Roscoff, UMR8227 (CNRS-Sorbonne Université) Dittami, Simon; Station Biologique de Roscoff, UMR8227 (CNRS-Sorbonne Université) Brunet, Maéva; Station Biologique de Roscoff, UMR8227 (CNRS-Sorbonne Université) Le Duff, Nolwen; Station Biologique de Roscoff, UMR8227 (CNRS-Sorbonne Université) Tanguy, Gwenn; Station Biologique de Roscoff, FR2424, Genomer sequencing platform Leblanc, Catherine; Station Biologique de Roscoff, LBI2M, UMR8227, CNRS-UPMC Gobet , Angélique ; MARBEC
Keywords:	chloroplast, plastid, 16S rDNA, alga-associated bacteria, metabarcoding, V34 primer, algal holobiont

Evaluation of a new primer combination to minimize plastid contamination in 16S rDNA metabarcoding analyses of alga-associated bacterial communities

François Thomas[1*], Simon M. Dittami[1], Maéva Brunet[1], Nolwen Le Duff[1], Gwenn Tanguy[2], Catherine Leblanc[1] and Angélique Gobet[1,3*]

[1] Sorbonne Université, CNRS, Integrative Biology of Marine Models (LBI2M), Station Biologique de Roscoff (SBR), 29680 Roscoff, France

[2] CNRS, Sorbonne Université, FR2424, Genomer, Station Biologique de Roscoff, 29680 Roscoff, France

[3] MARBEC, Ifremer, IRD, Université de Montpellier, CNRS, 34203 Sète, France

* Corresponding authors:

François Thomas
Marine Glycobiology group, UMR8227
Station Biologique de Roscoff
Place George Teissier
29680 Roscoff, France
email: fthomas@sb-roscoff.fr
Ph: +33256452148
Fax: +33298292324

Angélique Gobet
IFREMER – UMR MARBEC
Avenue Jean Monnet CS 30171
34203 Sète, France
email: angelique.gobet@ifremer.fr
Ph: +33499573250
Fax: +33499573294

Keywords

Chloroplast, plastid, 16S rDNA, alga-associated bacteria, metabarcoding, V34 primer, algal holobiont

Running title: Avoiding plastids in metabarcoding of algal microbiota

Originality-Significance statement

- Our aim was to minimize the amplification of plastidial 16S rDNA when analyzing alga-associated bacterial communities using metabarcoding.
- We compared the predicted coverage of five primer combinations *in silico*, including a universal primer set and four new sets designed to avoid amplification from plastids.
- Sequencing of a synthetic community showed similar performances for the universal primer set and one new combination.
- On natural alga-associated microbiota, the relative abundance of plastid sequences was reduced by 99-100% with the new primer set while the recovered bacterial richness was higher than with the universal primers given identical sequencing efforts and the community structure was maintained.

Summary

Plant- and alga-associated bacterial communities are generally described via 16S rDNA metabarcoding using universal primers. As plastid genomes encode 16S rDNA related to cyanobacteria, these datasets frequently contain >90% plastidial sequences, and the bacterial diversity may be under-sampled. To overcome this limitation we evaluated *in silico* the taxonomic coverage for four primer combinations targeting the 16S rDNA V3-V4 region. They included a forward primer universal to *Bacteria* (S-D-Bact-0341-b-S-17) and four reverse primers designed to avoid plastid DNA amplification. The best primer combination

(NOCHL) was compared to the universal primer set in the wet lab using a synthetic community and samples from 3 macroalgal species. The proportion of plastid sequences was reduced by 99-100% with the NOCHL primers compared to the universal primers, irrespective of algal hosts, sample collection and extraction protocols. Additionally, the NOCHL primers yielded a higher richness while maintaining the community structure. As *Planctomycetes*, *Verrucomicrobia*, and *Cyanobacteria* were underrepresented (70-90%) compared to universal primers, combining the NOCHL set with taxon-specific primers may be useful for a complete description of the alga-associated bacterial diversity. The NOCHL primers represent an innovation to study algal holobionts without amplifying host plastid sequences and may further be applied on other photosynthetic hosts.

Introduction

Numerous studies on algae-associated bacteria have highlighted their importance for health and physiology of the algal holobiont as well as for biomass degradation (Barott *et al.*, 2011; Egan *et al.*, 2013, 2014; Singh and Reddy, 2016). Bacteria can be abundant on macroalgal surfaces, ranging from 10^6 to 10^7 cells per cm^2, and encompass a large phylogenetic diversity (van der Loos *et al.*, 2019). These communities are distinct from bacterial assemblages found in the surrounding seawater or on inert surfaces (Burke *et al.*, 2011; Stratil *et al.*, 2014; Lemay *et al.*, 2018). The composition of algae-associated bacterial communities has been shown to be host-specific and to vary depending on algal tissue, seasons, sampling sites and physiological status of the algae (Staufenberger *et al.*, 2008; Lachnit *et al.*, 2011; Miranda *et al.*, 2013; Stratil *et al.*, 2013; Zozaya-Valdes *et al.*, 2015; Aires *et al.*, 2016; Paix *et al.*, 2019). A common approach to characterize these algae-associated communities is metabarcoding, whereby a selected variable region of the 16S rDNA is sequenced using a primer set universal to *Bacteria* (e.g., forward: S-D-Bact-0341-b-S-17 / reverse: S-D-Bact-0785-a-A-21

(Klindworth et al, 2013)) Due to the classical sampling strategies for macroalgae-associated microbiomes (e.g. algal surface scraping or tissue grinding (Bengtsson *et al.*, 2012; Aires *et al.*, 2016)), one major issue is that bacterial DNA samples are often contaminated with algal DNA. Primers with a large taxonomic coverage for *Bacteria* will tend to also amplify 16S rDNA sequences from plastids. This is because plastid genomes, deriving from endosymbiosis events, still encode 16S rDNA closely related to their cyanobacterial ancestors. Datasets obtained from brown algal field samples using common universal primers may contain over 90% of sequences affiliated to plastid DNA (Leblanc, personal data). This high proportion of plastidial sequences decreases the sequencing depth for bacterial epibionts and reduces the power of diversity analyses. One strategy to allow a large coverage of *Bacteria* while minimizing 16S gene amplification from plastids is to design a primer in a region that differs between plastid sequences and those of *Bacteria*. Chelius and Triplett have identified such a region between the positions 783 and 799 of the 16S rDNA (following the *E. coli* numbering system) and they designed the 799F primer which includes four mismatches with chloroplasts to amplify *Bacteria* while avoiding chloroplast amplification from maize roots (Chelius and Triplett, 2001). The resulting primer was then used in combination with primer 1193R (amplification product 394 bp) in several studies on bacterial communities associated with plants (Sagaram *et al.*, 2009; Bodenhausen *et al.*, 2013) or macroalgae (Vieira *et al.*, 2016; Aires *et al.*, 2018; Serebryakova *et al.*, 2018). However, this primer set still yielded up to 38% of sequences affiliated to chloroplasts (Sagaram *et al.*, 2009). Several sets of primers including the 799F primer with or without modifications were subsequently designed and tested to minimize plastid contamination on samples from plants and algae (Hanshew *et al.*, 2013; Miranda *et al.*, 2013; Aires *et al.*, 2016). These primer sets target the V5-V8 or V5-V9 regions, yielding products of ca. 590 bp or 750 bp, respectively. Although suitable for the 454 pyrosequencing technology used in the latter studies, this fragment length

is not compatible with the current Illumina v3 sequencing chemistry, which produces paired-end reads of 2 × 300 bp. Allowing for a recommended overlap of at least 50 bp, the total product length for this technology should not exceed 500 bp. Our objective was to evaluate primer combinations matching these criteria while minimizing plastid contamination in 16S rDNA metabarcoding analyses by: (i) comparing the efficiency of a primer set universal to *Bacteria* and that of four primer combinations avoiding plastid amplification *in silico*, (ii) comparing the efficiency of the best primer combination from the *in silico* results with that of the universal primer set in the wet lab.

Results and Discussion

In silico evaluation of primer combinations to minimize the amplification of plastid sequences

We evaluated several sets of primers to amplify a 450 bp fragment spanning the V3 and V4 regions of the 16S rDNA while minimizing amplification of plastid sequences (Table S1). These primer sets were designed by combining the universal forward primer S-D-Bact-0341-b-S-17 (Klindworth *et al.*, 2013) with reverse primers corresponding to the reverse complement of four V5-V8 forward primers previously used in 454 pyrosequencing to minimize chloroplast contamination (Hanshew *et al.*, 2013). The four new primer combinations (*E. coli* position 341 to 785) covers most of the region amplified by the original V34 set (*E. coli* position 341 to 799), ensuring the comparability of data. The performance of these different primer combinations was tested by running an *in silico* PCR on the SILVA database ssu-132 with the RefNR sequence collection, using the online tool SILVA TestPrime 1.0 (Klindworth *et al.*, 2013). The universal V34 combination with the original reverse primer S-D-Bact-0785-a-A-21 had an *in silico* predicted coverage for plastids (Table 1) ranging from 57.3% to 88.3% with zero or 2 allowed mismatches, respectively. By contrast, three of the

four new combinations (NOCHL, NOCHL3, and NOCHL6 but not NOCHL7, Table S1) showed consistently low predicted coverage for plastids, even with 2 allowed mismatches (2.7%, 1.1% and 1.1%, respectively). The overall coverage of these three primer combinations for all bacterial sequences remained high (79% with no mismatch, 85-90% with two mismatches). However, the NOCHL3 and NOCHL6 primer sets were predicted to perform poorly on a number of phyla, including *Planctomycetes* (12% coverage with two mismatches) and *Verrucomicrobia* (17% coverage), which are known to be part of alga-associated microbial communities (Bengtsson and Øvreås, 2010; Lage and Bondoso, 2014; Vollmers *et al.*, 2017). For all taxa, the NOCHL combination had equal or better coverage than the NOCHL3 and NOCHL6 combinations, notably for *Planctomycetes* (82.3%) and *Verrucomicrobia* (48.7%). The NOCHL combination was therefore considered the most promising candidate to minimize plastid contamination while maintaining the overall bacterial diversity, and chosen to prepare Illumina-sequencing libraries.

Comparison of primer performances in the wet lab

Performances of the NOCHL and V34 primer combinations were compared *in vitro* using different types of samples. As a positive control for the metabarcoding experiment, we constructed a mock community based on the genomic DNA of pure bacterial isolates. Genomic DNA was extracted from a total of 32 sequenced bacterial strains (Table S2) and mixed in known proportions to assemble the mock community. Furthermore, the microbiota associated with three brown algal species was sampled according to different approaches. The filamentous brown alga *Ectocarpus subulatus* samples were collected 5 km North of the Hopkins River Falls (Australia) and whole tissues were ground before DNA extraction. Blade samples of *Laminaria hyperborea* sporophytes were collected at the coast off Roscoff (Brittany, France) and soaked in a lysis buffer for DNA extraction. The blades of *L. digitata*

sporophytes from Roscoff were swabbed for DNA extraction. All samplings were performed in triplicate. Details on sampling and DNA extraction protocols are available in Supplementary Information. Amplicon and library preparation were carried out following the standard Illumina protocol (Illumina web page, 2018), including several negative controls, as detailed in Supplementary Information. Both the NOCHL and V34 primer pairs successfully amplified a ca. 450-bp fragment covering a similar region (V3-V4) of the bacterial 16S rDNA. Sequencing libraries were prepared in parallel for each sample (nine algal samples, one mock community, and six negative controls) using each primer pair. Sequencing was carried out using one run of an Illumina Miseq yielding a total of 5,549,008 read pairs. Sequences were deposited at the ENA under project accession number PRJEB33453. After quality checks and removal of exogenous contaminations, low-quality sequences, and blanks (see Supplementary Information), 2,784,253 assembled reads remained for data analysis. Sequences were clustered, assigned to operational taxonomic units (OTUs) at 97% identity, taxonomically classified with RDP classifier (Wang *et al.*, 2007) on the Silva SSU database release 132 (Quast *et al.*, 2013), and filtered to remove rare OTUs (see details in Supplementary Information). The final dataset comprised 3,009 OTUs.

First, we compared metabarcoding results from the primer combinations V34 and NOCHL on the mock community. The sequencing error rate calculated as in Kozich *et al.* (2013) from the mock community datasets was 0.54% for both the V34 and NOCHL libraries. A total of 65 OTUs were detected in the rarefied mock community datasets. The expected relative proportion p_i of 16S rDNAs from strain i was calculated as follows:

184 $$p_i = \frac{\frac{q_i}{M_i} \times c_i}{\sum_{i=1}^{92} \frac{q_i}{M_i} \times c_i}$$

185 where q_i is the mass of DNA from strain i added in the MOCK community, M_i is the
186 molecular weight of the complete genome for strain i and c_i is the number of 16S rDNA gene
187 copies in the genome of strain i. At the genus level, poor correlations (Pearson coefficient
188 <0.4) were obtained between the expected 16S rDNA proportions in the mock community
189 and the relative abundance of taxa found in V34 or NOCHL libraries (Table 2). From the 26
190 genera represented in the mock community, 18 and 20 were detected in the V34 and NOCHL
191 datasets, respectively. Both primer combinations did not detect the 6 genera *Nonlabens*,
192 *Agrococcus*, *Arthrobacter*, *Dokdonia*, *Roseovarius*, and *Imperialibacter*. Sequences from
193 these genera might either not be amplified by the primer sets, or more likely be merged to
194 closely related sequences during OTU clustering. In addition, *Roseobacter* and *Hoeflea* were
195 missed by the V34 primer combination, whereas their relative abundance was close to the
196 expected value in the NOCHL dataset. With both primer pairs, ca. 30% of sequences could
197 not be classified down to the genus level, leading to discrepancies between expected and
198 measured proportions. A striking example is *Vibrio*: its expected abundance of 24.63% was
199 largely underestimated at ca. 8% with both primer pairs, whereas 11% of sequences were
200 assigned to unclassified *Vibrionaceae*. On the contrary, both primer combinations largely
201 overestimated the proportion of *Pseudoalteromonas*, possibly due to the proximity with
202 sequences from *Alteromonas* (Gauthier *et al.*, 1995). Although the NOCHL combination had
203 a suboptimal predicted coverage for *Planctomycetes* (Table 1), it still detected the
204 planctomycetal genus *Rhodopirellula* in the mock community, underestimating it by 56%. On
205 the other hand, the V34 combination overestimated *Rhodopirellula* by 47%. As expected,
206 correlations between theoretical and measured proportions were much better (Pearson
207 coefficient > 0.9) at the family level for both primer combinations (Table S3). All 13 families

represented in the mock community were detected using the NOCHL combination, whereas the V34 combination missed the *Phyllobacteriaceae* family. Altogether, the performance of the NOCHL combination was found to be comparable to the V34 set on the mock community in terms of sensitivity and accuracy.

We further compared the efficiency of the primer combinations V34 and NOCHL in field samples, characterizing the bacterial community associated with *Ectocarpus subulatus* (ECTO), *Laminaria digitata* (LDIG) and *Laminaria hyperborea* (LHYP). The relative abundance of sequences affiliated to plastids ranged from 0.6% to 66% in the V34 dataset (Figure 1). The highest relative abundances of plastid sequences were found when algal tissues were ground (ECTO) or soaked in lysis buffer (LHYP) before DNA extraction, compared to the swab-based technique (LDIG). It was drastically reduced by 99-100% with the NOCHL combination in all three types of alga-associated samples analyzed (Figure 1), with relative abundance of plastid 16S rDNA sequences ranging from 0 to 0.3%. This confirms that the new primer combination succeeds at minimizing plastid contamination, irrespective of the algal host, sample collection and DNA extraction protocol. The NOCHL combination was more efficient at reducing plastid contamination compared to a previous Illumina metabarcoding analysis of *Ectocarpus*-associated bacterial communities. In that former study, a modified set of 341F and 806R primers still yielded 32% of plastid sequences despite a central mismatch to avoid plastid DNA amplification (Dittami *et al.*, 2016).

Alpha-diversity analyses tended to show higher values of OTU richness (observed OTUs) and diversity (Shannon and Simpson indices) with the NOCHL combination compared to V34 before removal of plastid sequences (Figure 2A). This overall effect was found significant for Shannon and Simpson diversity indices and was more pronounced for samples from *E. subulatus* and *L. hyperborea*, After removal of plastid sequences (Figure 2B), the NOCHL

primer combination yielded significantly higher richness than V34 (Student's t-test, P=0.05) with an overall increase of 10%. Again, this effect was more pronounced for ECTO and LHYP samples. Indeed, ECTO and LHYP samples, where plastid contamination was the highest (Figure 1), showed the strongest effect of plastidial sequence removal on sequencing depth (Figure S1). This exemplifies the issue faced with universal primers, where a large amount of plastid-affiliated sequences can decrease the sequencing depth for target bacterial sequences and reduce the evenness. After removal of plastid sequences, lower Shannon and Simpson diversity indices were detected with the NOCHL combination compared with V34 (-3.2% and -0.5% lower values for Shannon and Simpson, respectively). This was mostly due to an increase in these indices for V34 datasets after removal of plastid sequences, while they stayed stable for the NOCHL datasets. Although not the main focus of this work, we noticed that bacterial communities retrieved from *Ectocarpus* were more diverse than those from the two *Laminaria* species. This might be due to the nature of the samples as well as the extraction protocol where entire algal specimens were ground for *Ectocarpus* (i.e. including the endomicrobiota), while for *Laminaria* the extraction protocols only targeted surface-attached bacteria. Joint hierarchical clustering of OTU-level datasets for both primer pairs after removal of plastid sequences showed that samples grouped according to algal host rather than primer pair (Figure 3A).

The relative abundance of phyla was generally similar between paired sets of samples (Figure 3B). Notable exceptions were the *Planctomycetes*, *Verrucomicrobia*, and *Cyanobacteria* that had lower relative abundance in the NOCHL libraries compared to V34 (Table S4). We further searched for differential OTUs between pairs of samples sequenced with the V34 or NOCHL combination. This analysis was performed with the edgeR package (Robinson *et al.*, 2009) on the non-transformed dataset after removal of plastid sequences, and accounted for the paired design of the study, *i.e.* the same samples were sequenced with the two primer

combinations. The significance threshold was set at α=5% after Benjamini-Hochberg correction for multiple testing. Out of the 2,852 OTUs remaining after discarding plastids, the edgeR analysis detected 39 differential OTUs (i.e. 1.4%) between primer combinations (Table S5). All differential OTUs were more abundant in the V34-amplified dataset compared to NOCHL. The majority of them belonged to *Verrucomicrobiae* and *Planctomycetacia* (Figure 4), partly reflecting the lower predicted coverage of NOCHL for these taxa. The apparent lower abundance of *Plantomycetacia* OTUs using the NOCHL primer set could also be in part due to their over-estimation with the V34 combination, as shown in the mock community analysis (Table 2). The highest fold-change was found for OTU00082 belonging to *Cyanobacteria*. The lower abundance of *Cyanobacteria* was inevitable since plastids and extant cyanobacteria share a common ancestor and therefore have homologous 16S rDNA sequences (Giovannoni *et al.*, 1988; Delwiche, 1999).

Conclusion

Considering that plastid sequences can represent more than 90% of all sequences in one sample when using universal primer sets, the bacterial diversity in a sample may be underrepresented. The new primer set NOCHL is efficient to avoid amplifying plastid sequences from an algal host while identifying a significantly higher bacterial richness than with the universal primers V34, given identical sequencing efforts. This validates that fewer plastid sequences in the samples lead to a larger access to bacterial sequences thus recovering more of the bacterial diversity. However, as some bacterial groups may be underrepresented, this primer set may be combined with taxon-specific primers (e.g. for *Planctomycetes*) for a more complete coverage of *Bacteria*. Still, the data obtained with the NOCHL and universal V34 primer sets are comparable since the targeted regions are the same and patterns of the community structure are similar. Although validated here on macroalgal samples, this primer

set may be useful to characterize bacterial communities from other environments where plastid contamination can be an issue such as terrestrial plants, microalgae, or gut microbiota of herbivores.

Acknowledgments and funding

The authors have no conflict of interest to declare. We warmly thank Gaëlle Correc and Sylvie Rousvoal for help with sampling as well as Florian de Bettignies and Dominique Davoult for providing samples. We thank Frédérique Le Roux and Christian Jeanthon for providing bacterial strains. This work has benefited from the facilities of the Genomer platform and from the computational resources of the ABiMS bioinformatics platform (FR 2424, CNRS-Sorbonne Université, Roscoff), which are part of the Biogenouest core facility network.

This work has benefited from the support of the French Government via the National Research Agency investment expenditure program IDEALG (ANR-10-BTBR-04). FT, SD, and CL acknowledge support by the Centre National de la Recherche Scientifique (CNRS). AG acknowledges support by the Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER).

References

Aires, T., Muyzer, G., Serrão, E.A., and Engelen, A.H. (2018) Unraveling seaweeds bacteriomes. In, Charrier,B., Wichard,T., and Reddy,C.R.K. (eds), *Protocols for Macroalgae Research*. CRC Press, pp. 95–113.

Aires, T., Serrão, E.A., and Engelen, A.H. (2016) Host and environmental specificity in bacterial communities associated to two highly invasive marine species (genus *Asparagopsis*). *Front Microbiol* **7**: 1–14.

Barott, K.L., Rodriguez-Brito, B., Janouškovec, J., Marhaver, K.L., Smith, J.E., Keeling, P., and Rohwer, F.L. (2011) Microbial diversity associated with four functional groups of benthic reef algae and the reef-building coral *Montastraea annularis*. *Environ Microbiol* **13**: 1192–1204.

Bengtsson, M.M. and Øvreås, L. (2010) *Planctomycetes* dominate biofilms on surfaces of the kelp *Laminaria hyperborea*. *BMC Microbiol* **10**: 261.

Bengtsson, M.M., Sjøtun, K., Lanzén, A., and Øvreås, L. (2012) Bacterial diversity in relation to secondary production and succession on surfaces of the kelp *Laminaria hyperborea*. *ISME J* **6**: 2188–2198.

Bodenhausen, N., Horton, M.W., and Bergelson, J. (2013) Bacterial Communities Associated with the Leaves and the Roots of *Arabidopsis thaliana*. *PLoS One* **8**:.

Burke, C., Thomas, T., Lewis, M., Steinberg, P., and Kjelleberg, S. (2011) Composition, uniqueness and variability of the epiphytic bacterial community of the green alga *Ulva australis*. *ISME J* **5**: 590–600.

Chelius, M.K. and Triplett, E.W. (2001) The diversity of archaea and bacteria in association with the roots of *Zea mays* L. *Microb Ecol* **41**: 252–263.

Delwiche (1999) Tracing the thread of plastid diversity through the tapestry of life. *Am Nat* **154**: S164.

Dittami, S.M., Duboscq-Bidot, L., Perennou, M., Gobet, A., Corre, E., Boyen, C., and Tonon, T. (2016) Host-microbe interactions as a driver of acclimation to salinity gradients in brown algal cultures. *ISME J* **10**: 51–63.

Egan, S., Fernandes, N.D., Kumar, V., Gardiner, M., and Thomas, T. (2014) Bacterial pathogens, virulence mechanism and host defence in marine macroalgae. *Environ Microbiol* **16**: 925–938.

Egan, S., Harder, T., Burke, C., Steinberg, P., Kjelleberg, S., and Thomas, T. (2013) The

seaweed holobiont: Understanding seaweed-bacteria interactions. *FEMS Microbiol Rev* **37**: 462–476.

Gauthier, G., Gauthier, M., and Christen, R. (1995) Phylogenetic analysis of the genera *Alteromonas*, *Shewanella*, and *Moritella* using genes coding for small-subunit rRNA sequences and division of the genus *Alteromonas* into two genera, *Alteromonas* (emended) and *Pseudoalteromonas* gen. nov., and proposal of twelve new species combinations. *Int J Syst Bacteriol* **45**: 755–761.

Giovannoni, S.J., Turner, S., Olsen, G.J., Barns, S., Lane, D.J., and Pace, N.R. (1988) Evolutionary relationships among cyanobacteria and green chloroplasts. *J Bacteriol* **170**: 3584–3592.

Hanshew, A.S., Mason, C.J., Raffa, K.F., and Currie, C.R. (2013) Minimization of chloroplast contamination in 16S rRNA gene pyrosequencing of insect herbivore bacterial communities. *J Microbiol Methods* **95**: 149–155.

Illumina web page (2018) 16S Metagenomic Sequencing Library Preparation.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glöckner, F.O. (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41**: 1–11.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., and Schloss, P.D. (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* **79**: 5112–20.

Lachnit, T., Meske, D., Wahl, M., Harder, T., and Schmitz, R. (2011) Epibacterial community patterns on marine macroalgae are host-specific but temporally variable. *Environ Microbiol* **13**: 655–665.

Lage, O.M. and Bondoso, J. (2014) *Planctomycetes* and macroalgae, a striking association.

Front Microbiol **5**: 1–9.

Lemay, M., Martone, P., Keeling, P., Burt, J., Krumhansl, K., Sanders, R., and Parfrey, L. (2018) Sympatric kelp species share a large portion of their surface bacterial communities. *Environ Microbiol* **20**: 658–670.

van der Loos, L.M., Eriksson, B.K., and Falcão Salles, J. (2019) The macroalgal holobiont in a changing sea. *Trends Microbiol* **27**: 635–650.

Miranda, L.N., Hutchison, K., Grossman, A.R., and Brawley, S.H. (2013) Diversity and abundance of the bacterial community of the red macroalga *Porphyra umbilicalis*: Did bacterial farmers produce macroalgae? *PLoS One* **8**: e58269.

Paix, B., Othmani, A., Debroas, D., Culioli, G., and Briand, J.F. (2019) Temporal covariation of epibacterial community and surface metabolome in the Mediterranean seaweed holobiont *Taonia atomaria*. *Environ Microbiol*. doi: 10.1111/1462-2920.14617

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590-6.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2009) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.

Sagaram, U.S., Deangelis, K.M., Trivedi, P., Andersen, G.L., Lu, S.E., and Wang, N. (2009) Bacterial diversity analysis of huanglongbing pathogen-infected citrus, using phyloChip arrays and 16S rRNA gene clone library sequencing. *Appl Environ Microbiol* **75**: 1566–1574.

Serebryakova, A., Aires, T., Viard, F., Serrão, E.A., and Engelen, A.H. (2018) Summer shifts of bacterial communities associated with the invasive brown seaweed *Sargassum muticum* are location and tissue dependent. *PLoS One* **13**: 1–18.

383 Singh, R.P. and Reddy, C.R.K. (2016) Unraveling the functions of the macroalgal
384 microbiome. *Front Microbiol* **6**: 1–8.
385 Staufenberger, T., Thiel, V., Wiese, J., and Imhoff, J.F. (2008) Phylogenetic analysis of
386 bacteria associated with *Laminaria saccharina*. *FEMS Microbiol Ecol* **64**: 65–77.
387 Stratil, S.B., Neulinger, S.C., Knecht, H., Friedrichs, A.K., and Wahl, M. (2014) Salinity
388 affects compositional traits of epibacterial communities on the brown macroalga *Fucus*
389 *vesiculosus*. *FEMS Microbiol Ecol* **88**: 272–279.
390 Stratil, S.B., Neulinger, S.C., Knecht, H., Friedrichs, A.K., and Wahl, M. (2013)
391 Temperature-driven shifts in the epibiotic bacterial community composition of the brown
392 macroalga *Fucus vesiculosus*. *Microbiologyopen* **2**: 338–349.
393 Vieira, C., Engelen, A.H., Guentas, L., Aires, T., Houlbreque, F., Gaubert, J., et al. (2016)
394 Species specificity of bacteria associated to the brown seaweeds *Lobophora* (Dictyotales,
395 Phaeophyceae) and their potential for induction of rapid coral bleaching in *Acropora*
396 *muricata*. *Front Microbiol* **7**: 1–13.
397 Vollmers, J., Frentrup, M., Rast, P., Jogler, C., and Kaster, A.K. (2017) Untangling genomes
398 of novel Planctomycetal and Verrucomicrobial species from Monterey bay kelp forest
399 metagenomes by refined binning. *Front Microbiol* **8**:.
400 Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007) Naive Bayesian classifier for
401 rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ*
402 *Microbiol* **73**: 5261–7.
403 Zozaya-Valdes, E., Egan, S., and Thomas, T. (2015) A comprehensive analysis of the
404 microbial communities of healthy and diseased marine macroalgae and the detection of
405 known and potential bacterial pathogens. *Front Microbiol* **6**: 1–9.
406

Table and Figure legends

Figure 1: Relative abundance of sequences affiliated to plastid OTUs using the V34 or NOCHL primer sets. Bacterial communities were analyzed in triplicate samples from different algal hosts: ECTO, *Ectocarpus subulatus*; LHYP, *Laminaria hyperborea*; LDIG, *Laminaria digitata*.

Figure 2: Richness and diversity estimates before (A) and after (B) removal of OTUs affiliated to plastids. Data were subsampled at 8,358 sequences per sample prior to analysis. Observed, observed OTU richness; Shannon, Shannon index; Simpson, Simpson index; ECTO, *Ectocarpus subulatus*; LDIG, *Laminaria digitata*; LHYP, *Laminaria hyperborea*. Results of paired Student's t-test testing the effect of the set of primers on each measured estimate are reported below each panel (df=8).

Figure 3: Effect of the primer pair on the detected bacterial community composition, after removal of OTUs affiliated to plastids.
(A) Hierarchical clustering calculated on Hellinger-transformed data based on Morisita-Horn distance using the complete linkage algorithm
(B) Taxonomic composition based on the phylum level

Figure 4: Taxonomic affiliation of the 39 differential OTUs detected with edgeR, depending on the primer pair. All of these OTUs received higher coverage with the V34 primers.

Supplementary Figure 1: Rarefaction analysis of datasets obtained with V34 and NOCHL primer sets before (A) and after (B) removal of plastid-related sequences.

Table 1: *In silico* predicted coverage of selected bacterial taxa, relative to the sequences available in the Silva SSU r132 database for different primer combinations using TestPrime 1.0. A length of 3 bases of 0-mismatch zone at 3' end of primers was chosen for tests with two allowed mismatches in the sequence.

Table 2: Genus-level comparison of theoretical relative abundance in the mock community with the measured abundance in libraries sequenced with V34 or NOCHL primers. unclass., unclassified.

Evaluation of a new primer combination to minimize plastid contamination in 16S rDNA metabarcoding analyses of alga-associated bacterial communities

François Thomas[1*], Simon M. Dittami[1], Maéva Brunet[1], Nolwen Le Duff[1], Gwenn Tanguy[2], Catherine Leblanc[1] and Angélique Gobet[1,3*]

[1] Sorbonne Université, CNRS, Integrative Biology of Marine Models (LBI2M), Station Biologique de Roscoff (SBR), 29680 Roscoff, France

[2] CNRS, Sorbonne Université, FR2424, Genomer, Station Biologique de Roscoff, 29680 Roscoff, France

[3] MARBEC, Ifremer, IRD, Université de Montpellier, CNRS, 34203 Sète, France

* Corresponding authors:

François Thomas
Marine Glycobiology group, UMR8227
Station Biologique de Roscoff
Place George Teissier
29680 Roscoff, France
email: fthomas@sb-roscoff.fr
Ph: +33256452148
Fax: +33298292324

Angélique Gobet
IFREMER – UMR MARBEC
Avenue Jean Monnet CS 30171
34203 Sète, France
email: angelique.gobet@ifremer.fr
Ph: +33499573250
Fax: +33499573294

Keywords

Chloroplast, plastid, 16S rDNA, alga-associated bacteria, metabarcoding, V34 primer, algal holobiont

Running title: Avoiding plastids in metabarcoding of algal microbiota

Originality-Significance statement

- Our aim was to minimize the amplification of plastidial 16S rDNA when analyzing alga-associated bacterial communities using metabarcoding.
- We compared the predicted coverage of five primer combinations *in silico*, including a universal primer set and four new sets designed to avoid amplification from plastids.
- Sequencing of a synthetic community showed similar performances for the universal primer set and one new combination.
- On natural alga-associated microbiota, the relative abundance of plastid sequences was reduced by 99-100% with the new primer set while the recovered bacterial richness was higher than with the universal primers given identical sequencing efforts and the community structure was maintained.

Summary

Plant- and alga-associated bacterial communities are generally described via 16S rDNA metabarcoding using universal primers. As plastid genomes encode 16S rDNA related to cyanobacteria, these datasets frequently contain >90% plastidial sequences, and the bacterial diversity may be under-sampled. To overcome this limitation we evaluated *in silico* the taxonomic coverage for four primer combinations targeting the 16S rDNA V3-V4 region. They included a forward primer universal to *Bacteria* (S-D-Bact-0341-b-S-17) and four reverse primers designed to avoid plastid DNA amplification. The best primer combination

(NOCHL) was compared to the universal primer set in the wet lab using a synthetic community and samples from 3 macroalgal species. The proportion of plastid sequences was reduced by 99-100% with the NOCHL primers compared to the universal primers, irrespective of algal hosts, sample collection and extraction protocols. Additionally, the NOCHL primers yielded a higher richness while maintaining the community structure. As *Planctomycetes*, *Verrucomicrobia*, and *Cyanobacteria* were underrepresented (70-90%) compared to universal primers, combining the NOCHL set with taxon-specific primers may be useful for a complete description of the alga-associated bacterial diversity. The NOCHL primers represent an innovation to study algal holobionts without amplifying host plastid sequences and may further be applied on other photosynthetic hosts.

Introduction

Numerous studies on algae-associated bacteria have highlighted their importance for health and physiology of the algal holobiont as well as for biomass degradation (Barott *et al.*, 2011; Egan *et al.*, 2013, 2014; Singh and Reddy, 2016). Bacteria can be abundant on macroalgal surfaces, ranging from 10^6 to 10^7 cells per cm^2, and encompass a large phylogenetic diversity (van der Loos *et al.*, 2019). These communities are distinct from bacterial assemblages found in the surrounding seawater or on inert surfaces (Burke *et al.*, 2011; Stratil *et al.*, 2014; Lemay *et al.*, 2018). The composition of algae-associated bacterial communities has been shown to be host-specific and to vary depending on algal tissue, seasons, sampling sites and physiological status of the algae (Staufenberger *et al.*, 2008; Lachnit *et al.*, 2011; Miranda *et al.*, 2013; Stratil *et al.*, 2013; Zozaya-Valdes *et al.*, 2015; Aires *et al.*, 2016; Paix *et al.*, 2019). A common approach to characterize these algae-associated communities is metabarcoding, whereby a selected variable region of the 16S rDNA is sequenced using a primer set universal to *Bacteria* (e.g., forward: S-D-Bact-0341-b-S-17 / reverse: S-D-Bact-0785-a-A-21

(Klindworth et al, 2013)) Due to the classical sampling strategies for macroalgae-associated microbiomes (e.g. algal surface scraping or tissue grinding (Bengtsson *et al.*, 2012; Aires *et al.*, 2016)), one major issue is that bacterial DNA samples are often contaminated with algal DNA. Primers with a large taxonomic coverage for *Bacteria* will tend to also amplify 16S rDNA sequences from plastids. This is because plastid genomes, deriving from endosymbiosis events, still encode 16S rDNA closely related to their cyanobacterial ancestors. Datasets obtained from brown algal field samples using common universal primers may contain over 90% of sequences affiliated to plastid DNA (Leblanc, personal data). This high proportion of plastidial sequences decreases the sequencing depth for bacterial epibionts and reduces the power of diversity analyses. One strategy to allow a large coverage of *Bacteria* while minimizing 16S gene amplification from plastids is to design a primer in a region that differs between plastid sequences and those of *Bacteria*. Chelius and Triplett have identified such a region between the positions 783 and 799 of the 16S rDNA (following the *E. coli* numbering system) and they designed the 799F primer which includes four mismatches with chloroplasts to amplify *Bacteria* while avoiding chloroplast amplification from maize roots (Chelius and Triplett, 2001). The resulting primer was then used in combination with primer 1193R (amplification product 394 bp) in several studies on bacterial communities associated with plants (Sagaram *et al.*, 2009; Bodenhausen *et al.*, 2013) or macroalgae (Vieira *et al.*, 2016; Aires *et al.*, 2018; Serebryakova *et al.*, 2018). However, this primer set still yielded up to 38% of sequences affiliated to chloroplasts (Sagaram *et al.*, 2009). Several sets of primers including the 799F primer with or without modifications were subsequently designed and tested to minimize plastid contamination on samples from plants and algae (Hanshew *et al.*, 2013; Miranda *et al.*, 2013; Aires *et al.*, 2016). These primer sets target the V5-V8 or V5-V9 regions, yielding products of ca. 590 bp or 750 bp, respectively. Although suitable for the 454 pyrosequencing technology used in the latter studies, this fragment length

is not compatible with the current Illumina v3 sequencing chemistry, which produces paired-end reads of 2 × 300 bp. Allowing for a recommended overlap of at least 50 bp, the total product length for this technology should not exceed 500 bp. Our objective was to evaluate primer combinations matching these criteria while minimizing plastid contamination in 16S rDNA metabarcoding analyses by: (i) comparing the efficiency of a primer set universal to *Bacteria* and that of four primer combinations avoiding plastid amplification *in silico*, (ii) comparing the efficiency of the best primer combination from the *in silico* results with that of the universal primer set in the wet lab.

Results and Discussion

In silico evaluation of primer combinations to minimize the amplification of plastid sequences

We evaluated several sets of primers to amplify a 450 bp fragment spanning the V3 and V4 regions of the 16S rDNA while minimizing amplification of plastid sequences (Table S1). These primer sets were designed by combining the universal forward primer S-D-Bact-0341-b-S-17 (Klindworth *et al.*, 2013) with reverse primers corresponding to the reverse complement of four V5-V8 forward primers previously used in 454 pyrosequencing to minimize chloroplast contamination (Hanshew *et al.*, 2013). The four new primer combinations (*E. coli* position 341 to 785) covers most of the region amplified by the original V34 set (*E. coli* position 341 to 799), ensuring the comparability of data. The performance of these different primer combinations was tested by running an *in silico* PCR on the SILVA database ssu-132 with the RefNR sequence collection, using the online tool SILVA TestPrime 1.0 (Klindworth *et al.*, 2013). The universal V34 combination with the original reverse primer S-D-Bact-0785-a-A-21 had an *in silico* predicted coverage for plastids (Table 1) ranging from 57.3% to 88.3% with zero or 2 allowed mismatches, respectively. By contrast, three of the

four new combinations (NOCHL, NOCHL3, and NOCHL6 but not NOCHL7, Table S1) showed consistently low predicted coverage for plastids, even with 2 allowed mismatches (2.7%, 1.1% and 1.1%, respectively). The overall coverage of these three primer combinations for all bacterial sequences remained high (79% with no mismatch, 85-90% with two mismatches). However, the NOCHL3 and NOCHL6 primer sets were predicted to perform poorly on a number of phyla, including *Planctomycetes* (12% coverage with two mismatches) and *Verrucomicrobia* (17% coverage), which are known to be part of alga-associated microbial communities (Bengtsson and Øvreås, 2010; Lage and Bondoso, 2014; Vollmers *et al.*, 2017). For all taxa, the NOCHL combination had equal or better coverage than the NOCHL3 and NOCHL6 combinations, notably for *Planctomycetes* (82.3%) and *Verrucomicrobia* (48.7%). The NOCHL combination was therefore considered the most promising candidate to minimize plastid contamination while maintaining the overall bacterial diversity, and chosen to prepare Illumina-sequencing libraries.

Comparison of primer performances in the wet lab

Performances of the NOCHL and V34 primer combinations were compared *in vitro* using different types of samples. As a positive control for the metabarcoding experiment, we constructed a mock community based on the genomic DNA of pure bacterial isolates. Genomic DNA was extracted from a total of 32 sequenced bacterial strains (Table S2) and mixed in known proportions to assemble the mock community. Furthermore, the microbiota associated with three brown algal species was sampled according to different approaches. The filamentous brown alga *Ectocarpus subulatus* samples were collected 5 km North of the Hopkins River Falls (Australia) and whole tissues were ground before DNA extraction. Blade samples of *Laminaria hyperborea* sporophytes were collected at the coast off Roscoff (Brittany, France) and soaked in a lysis buffer for DNA extraction. The blades of *L. digitata*

sporophytes from Roscoff were swabbed for DNA extraction. All samplings were performed in triplicate. Details on sampling and DNA extraction protocols are available in Supplementary Information. Amplicon and library preparation were carried out following the standard Illumina protocol (Illumina web page, 2018), including several negative controls, as detailed in Supplementary Information. Both the NOCHL and V34 primer pairs successfully amplified a ca. 450-bp fragment covering a similar region (V3-V4) of the bacterial 16S rDNA. Sequencing libraries were prepared in parallel for each sample (nine algal samples, one mock community, and six negative controls) using each primer pair. Sequencing was carried out using one run of an Illumina Miseq yielding a total of 5,549,008 read pairs. Sequences were deposited at the ENA under project accession number PRJEB33453. After quality checks and removal of exogenous contaminations, low-quality sequences, and blanks (see Supplementary Information), 2,784,253 assembled reads remained for data analysis. Sequences were clustered, assigned to operational taxonomic units (OTUs) at 97% identity, taxonomically classified with RDP classifier (Wang *et al.*, 2007) on the Silva SSU database release 132 (Quast *et al.*, 2013), and filtered to remove rare OTUs (see details in Supplementary Information). The final dataset comprised 3,009 OTUs.

First, we compared metabarcoding results from the primer combinations V34 and NOCHL on the mock community. The sequencing error rate calculated as in Kozich *et al.* (2013) from the mock community datasets was 0.54% for both the V34 and NOCHL libraries. A total of 65 OTUs were detected in the rarefied mock community datasets. The expected relative proportion p_i of 16S rDNAs from strain i was calculated as follows:

$$p_i = \frac{\frac{q_i}{M_i} \times c_i}{\sum_{i=1}^{92} \frac{q_i}{M_i} \times c_i}$$

where q_i is the mass of DNA from strain *i* added in the MOCK community, M_i is the molecular weight of the complete genome for strain *i* and c_i is the number of 16S rDNA gene copies in the genome of strain *i*. At the genus level, poor correlations (Pearson coefficient <0.4) were obtained between the expected 16S rDNA proportions in the mock community and the relative abundance of taxa found in V34 or NOCHL libraries (Table 2). From the 26 genera represented in the mock community, 18 and 20 were detected in the V34 and NOCHL datasets, respectively. Both primer combinations did not detect the 6 genera *Nonlabens*, *Agrococcus*, *Arthrobacter*, *Dokdonia*, *Roseovarius*, and *Imperialibacter*. Sequences from these genera might either not be amplified by the primer sets, or more likely be merged to closely related sequences during OTU clustering. In addition, *Roseobacter* and *Hoeflea* were missed by the V34 primer combination, whereas their relative abundance was close to the expected value in the NOCHL dataset. With both primer pairs, ca. 30% of sequences could not be classified down to the genus level, leading to discrepancies between expected and measured proportions. A striking example is *Vibrio*: its expected abundance of 24.63% was largely underestimated at ca. 8% with both primer pairs, whereas 11% of sequences were assigned to unclassified *Vibrionaceae*. On the contrary, both primer combinations largely overestimated the proportion of *Pseudoalteromonas*, possibly due to the proximity with sequences from *Alteromonas* (Gauthier *et al.*, 1995). Although the NOCHL combination had a suboptimal predicted coverage for *Planctomycetes* (Table 1), it still detected the planctomycetal genus *Rhodopirellula* in the mock community, underestimating it by 56%. On the other hand, the V34 combination overestimated *Rhodopirellula* by 47%. As expected, correlations between theoretical and measured proportions were much better (Pearson coefficient > 0.9) at the family level for both primer combinations (Table S3). All 13 families

represented in the mock community were detected using the NOCHL combination, whereas the V34 combination missed the *Phyllobacteriaceae* family. Altogether, the performance of the NOCHL combination was found to be comparable to the V34 set on the mock community in terms of sensitivity and accuracy.

We further compared the efficiency of the primer combinations V34 and NOCHL in field samples, characterizing the bacterial community associated with *Ectocarpus subulatus* (ECTO), *Laminaria digitata* (LDIG) and *Laminaria hyperborea* (LHYP). The relative abundance of sequences affiliated to plastids ranged from 0.6% to 66% in the V34 dataset (Figure 1). The highest relative abundances of plastid sequences were found when algal tissues were ground (ECTO) or soaked in lysis buffer (LHYP) before DNA extraction, compared to the swab-based technique (LDIG). It was drastically reduced by 99-100% with the NOCHL combination in all three types of alga-associated samples analyzed (Figure 1), with relative abundance of plastid 16S rDNA sequences ranging from 0 to 0.3%. This confirms that the new primer combination succeeds at minimizing plastid contamination, irrespective of the algal host, sample collection and DNA extraction protocol. The NOCHL combination was more efficient at reducing plastid contamination compared to a previous Illumina metabarcoding analysis of *Ectocarpus*-associated bacterial communities. In that former study, a modified set of 341F and 806R primers still yielded 32% of plastid sequences despite a central mismatch to avoid plastid DNA amplification (Dittami *et al.*, 2016).

Alpha-diversity analyses tended to show higher values of OTU richness (observed OTUs ~~and Chao1 richness estimator~~) and diversity (Shannon and Simpson indices) with the NOCHL combination compared to V34 before removal of plastid sequences (Figure 2A). This overall effect was found significant for Shannon and Simpson diversity indices and was more pronounced for samples from *E. subulatus* and *L. hyperborea*, After removal of plastid

sequences (Figure 2B), the NOCHL primer combination yielded significantly higher richness estimates than V34 (Student's t-test, P=0.05) with an overall increase of 10%. Again, this effect was more pronounced for ECTO and LHYP samples. Indeed, ECTO and LHYP samples, where plastid contamination was the highest (Figure 1), showed the strongest effect of plastidial sequence removal on sequencing depth (Figure S1). This exemplifies the issue faced with universal primers, where a large amount of plastid-affiliated sequences can decrease the sequencing depth for target bacterial sequences and reduce the evenness. After removal of plastid sequences, lower Shannon and Simpson diversity indices were detected with the NOCHL combination compared with V34 (-3.2% and -0.5% lower values for Shannon and Simpson, respectively). This was mostly due to an increase in these indices for V34 datasets after removal of plastid sequences, while they stayed stable for the NOCHL datasets. Although not the main focus of this work, we noticed that bacterial communities retrieved from *Ectocarpus* were more diverse than those from the two *Laminaria* species. This might be due to the nature of the samples as well as the extraction protocol where entire algal specimens were ground for *Ectocarpus* (i.e. including the endomicrobiota), while for *Laminaria* the extraction protocols only targeted surface-attached bacteria. Joint hierarchical clustering of OTU-level datasets for both primer pairs after removal of plastid sequences showed that samples grouped according to algal host rather than primer pair (Figure 3A).

The relative abundance of phyla was generally similar between paired sets of samples (Figure 3B). Notable exceptions were the *Planctomycetes*, *Verrucomicrobia*, and *Cyanobacteria* that had lower relative abundance in the NOCHL libraries compared to V34 (Table S4). We further searched for differential OTUs between pairs of samples sequenced with the V34 or NOCHL combination. This analysis was performed with the edgeR package (Robinson *et al.*, 2009) on the non-transformed dataset after removal of plastid sequences, and accounted for the paired design of the study, *i.e.* the same samples were sequenced with the two primer

combinations. The significance threshold was set at α=5% after Benjamini-Hochberg correction for multiple testing. Out of the 2,852 OTUs remaining after discarding plastids, the edgeR analysis detected 39 differential OTUs (i.e. 1.4%) between primer combinations (Table S5). All differential OTUs were more abundant in the V34-amplified dataset compared to NOCHL. The majority of them belonged to *Verrucomicrobiae* and *Planctomycetacia* (Figure 4), partly reflecting the lower predicted coverage of NOCHL for these taxa. The apparent lower abundance of *Plantomycetacia* OTUs using the NOCHL primer set could also be in part due to their over-estimation with the V34 combination, as shown in the mock community analysis (Table 2). The highest fold-change was found for OTU00082 belonging to *Cyanobacteria*. The lower abundance of *Cyanobacteria* was inevitable since plastids and extant cyanobacteria share a common ancestor and therefore have homologous 16S rDNA sequences (Giovannoni *et al.*, 1988; Delwiche, 1999).

Conclusion

Considering that plastid sequences can represent more than 90% of all sequences in one sample when using universal primer sets, the bacterial diversity in a sample may be underrepresented. The new primer set NOCHL is efficient to avoid amplifying plastid sequences from an algal host while identifying a significantly higher bacterial richness than with the universal primers V34, given identical sequencing efforts. This validates that fewer plastid sequences in the samples lead to a larger access to bacterial sequences thus recovering more of the bacterial diversity. However, as some bacterial groups may be underrepresented, this primer set may be combined with taxon-specific primers (e.g. for *Planctomycetes*) for a more complete coverage of *Bacteria*. Still, the data obtained with the NOCHL and universal V34 primer sets are comparable since the targeted regions are the same and patterns of the community structure are similar. Although validated here on macroalgal samples, this primer

set may be useful to characterize bacterial communities from other environments where plastid contamination can be an issue such as terrestrial plants, microalgae, or gut microbiota of herbivores.

Acknowledgments and funding

The authors have no conflict of interest to declare. We warmly thank Gaëlle Correc and Sylvie Rousvoal for help with sampling as well as Florian de Bettignies and Dominique Davoult for providing samples. We thank Frédérique Le Roux and Christian Jeanthon for providing bacterial strains. This work has benefited from the facilities of the Genomer platform and from the computational resources of the ABiMS bioinformatics platform (FR 2424, CNRS-Sorbonne Université, Roscoff), which are part of the Biogenouest core facility network.

This work has benefited from the support of the French Government via the National Research Agency investment expenditure program IDEALG (ANR-10-BTBR-04). FT, SD, and CL acknowledge support by the Centre National de la Recherche Scientifique (CNRS). AG acknowledges support by the Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER).

References

Aires, T., Muyzer, G., Serrão, E.A., and Engelen, A.H. (2018) Unraveling seaweeds bacteriomes. In, Charrier,B., Wichard,T., and Reddy,C.R.K. (eds), *Protocols for Macroalgae Research*. CRC Press, pp. 95–113.

Aires, T., Serrão, E.A., and Engelen, A.H. (2016) Host and environmental specificity in bacterial communities associated to two highly invasive marine species (genus *Asparagopsis*). *Front Microbiol* **7**: 1–14.

Barott, K.L., Rodriguez-Brito, B., Janouškovec, J., Marhaver, K.L., Smith, J.E., Keeling, P., and Rohwer, F.L. (2011) Microbial diversity associated with four functional groups of benthic reef algae and the reef-building coral *Montastraea annularis*. *Environ Microbiol* **13**: 1192–1204.

Bengtsson, M.M. and Øvreås, L. (2010) *Planctomycetes* dominate biofilms on surfaces of the kelp *Laminaria hyperborea*. *BMC Microbiol* **10**: 261.

Bengtsson, M.M., Sjøtun, K., Lanzén, A., and Øvreås, L. (2012) Bacterial diversity in relation to secondary production and succession on surfaces of the kelp *Laminaria hyperborea*. *ISME J* **6**: 2188–2198.

Bodenhausen, N., Horton, M.W., and Bergelson, J. (2013) Bacterial Communities Associated with the Leaves and the Roots of *Arabidopsis thaliana*. *PLoS One* **8**:.

Burke, C., Thomas, T., Lewis, M., Steinberg, P., and Kjelleberg, S. (2011) Composition, uniqueness and variability of the epiphytic bacterial community of the green alga *Ulva australis*. *ISME J* **5**: 590–600.

Chelius, M.K. and Triplett, E.W. (2001) The diversity of archaea and bacteria in association with the roots of *Zea mays* L. *Microb Ecol* **41**: 252–263.

Delwiche (1999) Tracing the thread of plastid diversity through the tapestry of life. *Am Nat* **154**: S164.

Dittami, S.M., Duboscq-Bidot, L., Perennou, M., Gobet, A., Corre, E., Boyen, C., and Tonon, T. (2016) Host-microbe interactions as a driver of acclimation to salinity gradients in brown algal cultures. *ISME J* **10**: 51–63.

Egan, S., Fernandes, N.D., Kumar, V., Gardiner, M., and Thomas, T. (2014) Bacterial pathogens, virulence mechanism and host defence in marine macroalgae. *Environ Microbiol* **16**: 925–938.

Egan, S., Harder, T., Burke, C., Steinberg, P., Kjelleberg, S., and Thomas, T. (2013) The

seaweed holobiont: Understanding seaweed-bacteria interactions. *FEMS Microbiol Rev* **37**: 462–476.

Gauthier, G., Gauthier, M., and Christen, R. (1995) Phylogenetic analysis of the genera *Alteromonas*, *Shewanella*, and *Moritella* using genes coding for small-subunit rRNA sequences and division of the genus *Alteromonas* into two genera, *Alteromonas* (emended) and *Pseudoalteromonas* gen. nov., and proposal of twelve new species combinations. *Int J Syst Bacteriol* **45**: 755–761.

Giovannoni, S.J., Turner, S., Olsen, G.J., Barns, S., Lane, D.J., and Pace, N.R. (1988) Evolutionary relationships among cyanobacteria and green chloroplasts. *J Bacteriol* **170**: 3584–3592.

Hanshew, A.S., Mason, C.J., Raffa, K.F., and Currie, C.R. (2013) Minimization of chloroplast contamination in 16S rRNA gene pyrosequencing of insect herbivore bacterial communities. *J Microbiol Methods* **95**: 149–155.

Illumina web page (2018) 16S Metagenomic Sequencing Library Preparation.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glöckner, F.O. (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41**: 1–11.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., and Schloss, P.D. (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* **79**: 5112–20.

Lachnit, T., Meske, D., Wahl, M., Harder, T., and Schmitz, R. (2011) Epibacterial community patterns on marine macroalgae are host-specific but temporally variable. *Environ Microbiol* **13**: 655–665.

Lage, O.M. and Bondoso, J. (2014) *Planctomycetes* and macroalgae, a striking association.

Front Microbiol **5**: 1–9.

Lemay, M., Martone, P., Keeling, P., Burt, J., Krumhansl, K., Sanders, R., and Parfrey, L. (2018) Sympatric kelp species share a large portion of their surface bacterial communities. *Environ Microbiol* **20**: 658–670.

van der Loos, L.M., Eriksson, B.K., and Falcão Salles, J. (2019) The macroalgal holobiont in a changing sea. *Trends Microbiol* **27**: 635–650.

Miranda, L.N., Hutchison, K., Grossman, A.R., and Brawley, S.H. (2013) Diversity and abundance of the bacterial community of the red macroalga *Porphyra umbilicalis*: Did bacterial farmers produce macroalgae? *PLoS One* **8**: e58269.

Paix, B., Othmani, A., Debroas, D., Culioli, G., and Briand, J.F. (2019) Temporal covariation of epibacterial community and surface metabolome in the Mediterranean seaweed holobiont *Taonia atomaria*. *Environ Microbiol*. doi: 10.1111/1462-2920.14617

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590-6.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2009) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.

Sagaram, U.S., Deangelis, K.M., Trivedi, P., Andersen, G.L., Lu, S.E., and Wang, N. (2009) Bacterial diversity analysis of huanglongbing pathogen-infected citrus, using phyloChip arrays and 16S rRNA gene clone library sequencing. *Appl Environ Microbiol* **75**: 1566–1574.

Serebryakova, A., Aires, T., Viard, F., Serrão, E.A., and Engelen, A.H. (2018) Summer shifts of bacterial communities associated with the invasive brown seaweed *Sargassum muticum* are location and tissue dependent. *PLoS One* **13**: 1–18.

Singh, R.P. and Reddy, C.R.K. (2016) Unraveling the functions of the macroalgal microbiome. *Front Microbiol* **6**: 1–8.

Staufenberger, T., Thiel, V., Wiese, J., and Imhoff, J.F. (2008) Phylogenetic analysis of bacteria associated with *Laminaria saccharina*. *FEMS Microbiol Ecol* **64**: 65–77.

Stratil, S.B., Neulinger, S.C., Knecht, H., Friedrichs, A.K., and Wahl, M. (2014) Salinity affects compositional traits of epibacterial communities on the brown macroalga *Fucus vesiculosus*. *FEMS Microbiol Ecol* **88**: 272–279.

Stratil, S.B., Neulinger, S.C., Knecht, H., Friedrichs, A.K., and Wahl, M. (2013) Temperature-driven shifts in the epibiotic bacterial community composition of the brown macroalga *Fucus vesiculosus*. *Microbiologyopen* **2**: 338–349.

Vieira, C., Engelen, A.H., Guentas, L., Aires, T., Houlbreque, F., Gaubert, J., et al. (2016) Species specificity of bacteria associated to the brown seaweeds *Lobophora* (Dictyotales, Phaeophyceae) and their potential for induction of rapid coral bleaching in *Acropora muricata*. *Front Microbiol* **7**: 1–13.

Vollmers, J., Frentrup, M., Rast, P., Jogler, C., and Kaster, A.K. (2017) Untangling genomes of novel Planctomycetal and Verrucomicrobial species from Monterey bay kelp forest metagenomes by refined binning. *Front Microbiol* **8**:.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261–7.

Zozaya-Valdes, E., Egan, S., and Thomas, T. (2015) A comprehensive analysis of the microbial communities of healthy and diseased marine macroalgae and the detection of known and potential bacterial pathogens. *Front Microbiol* **6**: 1–9.

Table and Figure legends

Figure 1: Relative abundance of sequences affiliated to plastid OTUs using the V34 or NOCHL primer sets. Bacterial communities were analyzed in triplicate samples from different algal hosts: ECTO, *Ectocarpus subulatus*; LHYP, *Laminaria hyperborea*; LDIG, *Laminaria digitata*.

Figure 2: Richness and diversity estimates before (A) and after (B) removal of OTUs affiliated to plastids. Data were subsampled at 8,358 sequences per sample prior to analysis. Observed, observed OTU richness; ~~Chao1, Chao 1 estimated richness;~~ Shannon, Shannon index; Simpson, Simpson index; ECTO, *Ectocarpus subulatus*; LDIG, *Laminaria digitata*; LHYP, *Laminaria hyperborea*. Results of paired Student's t-test testing the effect of the set of primers on each measured estimate are reported below each panel (df=8).

Figure 3: Effect of the primer pair on the detected bacterial community composition, after removal of OTUs affiliated to plastids.
(A) Hierarchical clustering calculated on Hellinger-transformed data based on Morisita-Horn distance using the complete linkage algorithm
(B) Taxonomic composition based on the phylum level

Figure 4: Taxonomic affiliation of the 39 differential OTUs detected with edgeR, depending on the primer pair. All of these OTUs received higher coverage with the V34 primers.

Supplementary Figure 1: Rarefaction analysis of datasets obtained with V34 and NOCHL primer sets before (A) and after (B) removal of plastid-related sequences.

Table 1: *In silico* predicted coverage of selected bacterial taxa, relative to the sequences available in the Silva SSU r132 database for different primer combinations using TestPrime 1.0. A length of 3 bases of 0-mismatch zone at 3' end of primers was chosen for tests with two allowed mismatches in the sequence.

Table 2: Genus-level comparison of theoretical relative abundance in the mock community with the measured abundance in libraries sequenced with V34 or NOCHL primers. unclass., unclassified.

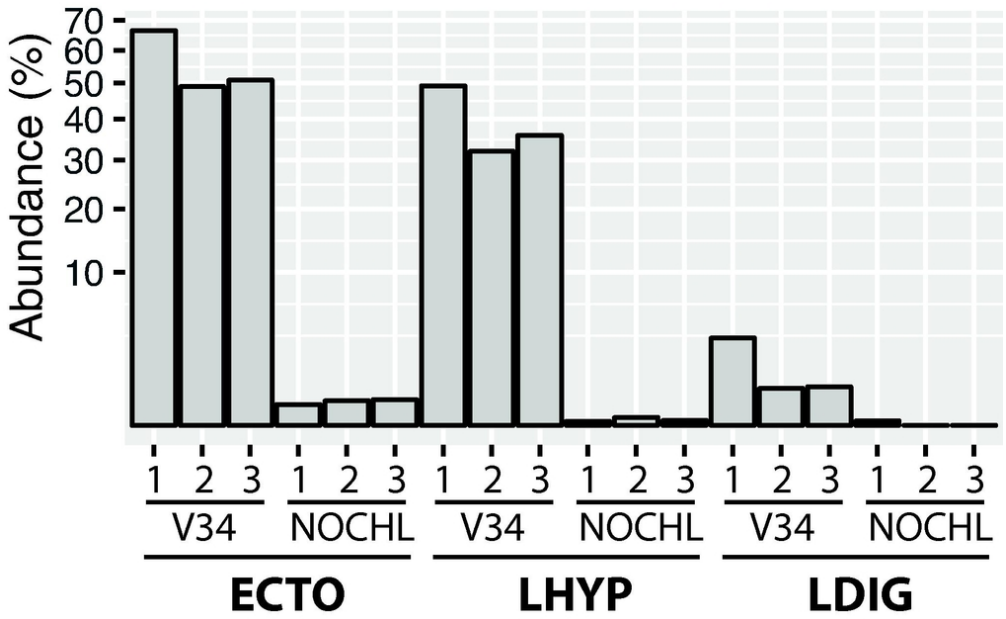

Figure 1: Relative abundance of sequences affiliated to plastid OTUs using the V34 or NOCHL primer sets. Bacterial communities were analyzed in triplicate samples from different algal hosts: ECTO, Ectocarpus subulatus; LHYP, Laminaria hyperborea; LDIG, Laminaria digitata.

85x53mm (300 x 300 DPI)

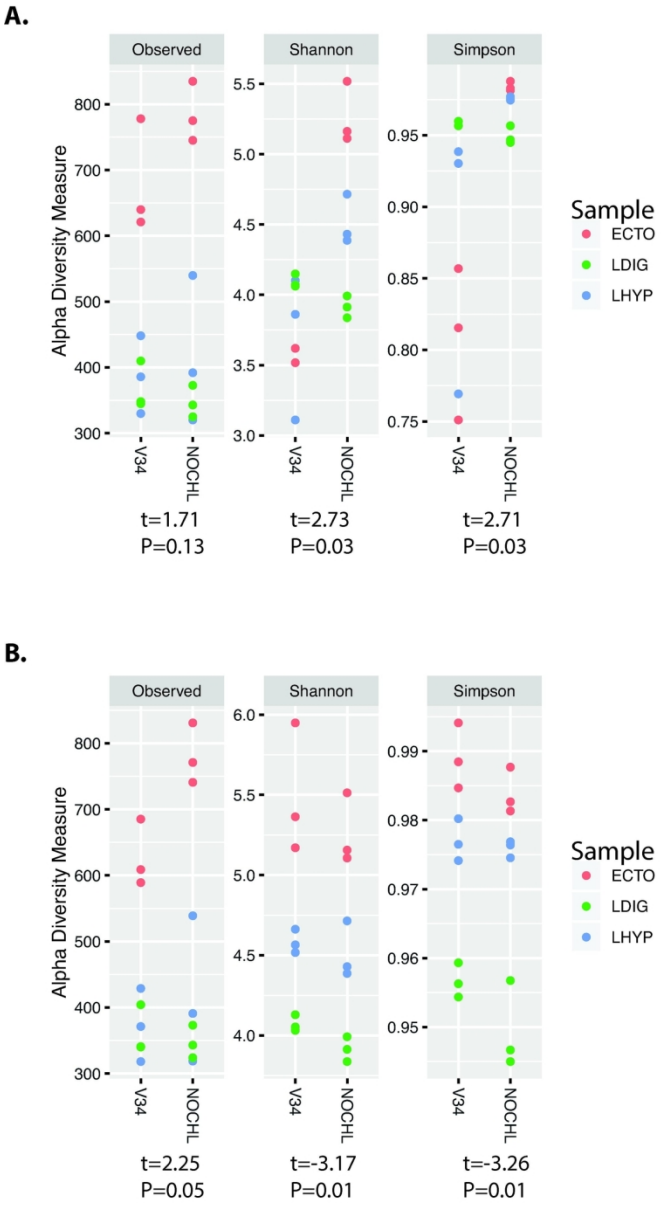

Figure 2: Richness and diversity estimates before (A) and after (B) removal of OTUs affiliated to plastids. Data were subsampled at 8,358 sequences per sample prior to analysis. Observed, observed OTU richness; Shannon, Shannon index; Simpson, Simpson index; ECTO, Ectocarpus subulatus; LDIG, Laminaria digitata; LHYP, Laminaria hyperborea. Results of paired Student's t-test testing the effect of the set of primers on each measured estimate are reported below each panel (df=8).

111x188mm (300 x 300 DPI)

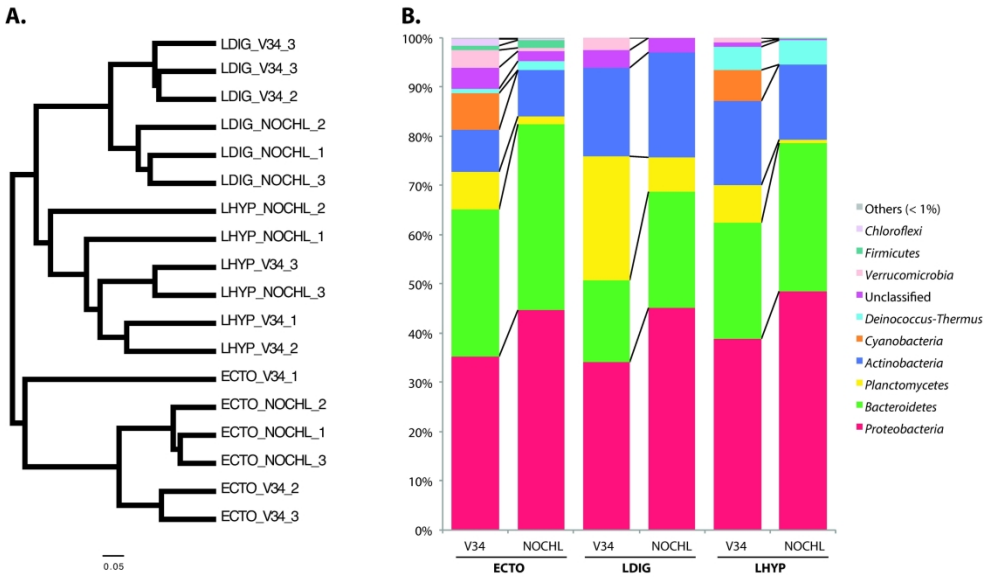

Figure 3: Effect of the primer pair on the detected bacterial community composition, after removal of OTUs affiliated to plastids.
(A) Hierarchical clustering calculated on Hellinger-transformed data based on Morisita-Horn distance using the complete linkage algorithm
(B) Taxonomic composition based on the phylum level

271x164mm (300 x 300 DPI)

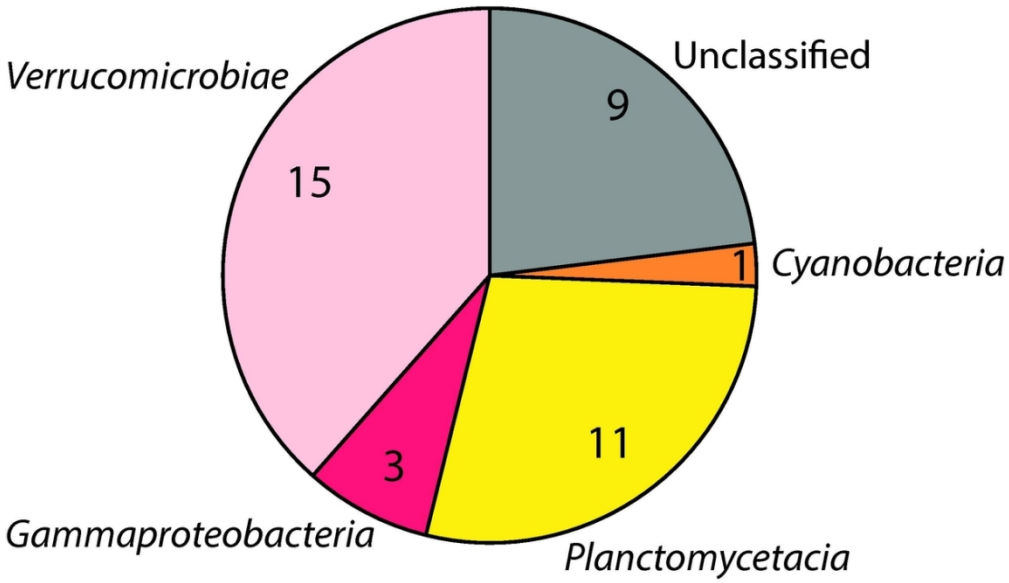

Figure 4: Taxonomic affiliation of the 39 differential OTUs detected with edgeR, depending on the primer pair

105x76mm (300 x 300 DPI)

Table 1: *In silico* predicted coverage of selected bacterial taxa, relative to the sequences available in the Silva SSU r132 database for different primer combinations using TestPrime 1.0. A length of 3 bases of 0-mismatch zone at 3' end of primers was chosen for tests with two allowed mismatches in the sequence.

	0 mismatch					2 mismatches				
	V34	NOCHL	NOCHL3	NOCHL6	NOCHL7	V34	NOCHL	NOCHL3	NOCHL6	NOCHL7
Acidobacteria	92.0	43.1	43.2	43.1	43.3	95.7	95.0	45.3	45.1	96.2
Actinobacteria	82.4	82.5	82.7	82.6	82.7	86.6	96.3	96.5	96.1	86.9
Aquificae	89.3	89.6	90.2	89.9	90.5	95.0	95.5	95.8	95.5	96.1
Bacteroidetes	89.8	88.8	89.3	89.0	89.4	95.4	95.8	95.2	94.8	95.9
Chlamydiae	82.8	65.9	65.9	65.9	66.4	96.7	96.9	79.7	79.7	96.7
Chloroflexi	39.0	19.3	20.8	20.7	21.1	88.7	48.6	30.3	30.3	88.6
Cyanobacteria	76.2	0.6	0.6	0.6	0.6	92.4	1.9	0.9	0.9	88.6
Chloroplasts	57.3	0.7	0.7	0.7	0.7	88.3	2.7	1.1	1.1	81.3
Deinococcus-Thermus	92.9	92.6	92.6	92.6	93.0	96.7	96.9	96.3	96.2	96.7
Epsilonbacteraeota	93.9	85.1	85.9	85.5	85.8	96.8	96.6	88.6	88.2	97.3
Firmicutes	88.2	84.0	84.9	84.7	85.0	94.2	92.8	90.5	90.1	94.7
Fusobacteria	87.1	86.1	87.0	86.7	87.2	94.9	94.0	93.9	93.6	95.3
Gemmatimonadetes	89.6	89.2	89.8	89.4	89.8	93.6	94.7	95.5	94.4	94.9
Planctomycetes	73.5	10.6	10.6	10.5	10.7	85.1	82.3	12.0	11.9	85.8
Proteobacteria	90.0	89.0	89.4	89.2	89.4	95.7	94.3	93.6	93.1	96.2
Spirochaetes	75.7	70.5	70.7	70.3	70.7	88.2	94.6	85.6	85.1	88.4
Verrucomicrobia	87.3	16.2	16.3	16.3	16.5	93.0	48.7	17.2	17.1	92.5
All Bacteria	**86.6**	**79.1**	**79.6**	**79.3**	**79.6**	**93.6**	**90.1**	**85.3**	**84.9**	**93.9**

Table 2: Comparison between the theoretical and the observed relative abundance of genera in the mock community in libraries sequenced with V34 or NOCHL primers. unclass., unclassified.

Genus	Relative abundance of sequences in the mock community (%)		
	Theory	Primers V34	Primers NOCHL
Vibrio	24.63	8.28	8.78
Formosa	9.72	3.04	2.53
Maribacter	9.42	5.41	5.24
Zobellia	7.24	5.07	3.04
Alteromonas	5.29	1.01	1.35
Nonlabens	5.02	0.00	0.00
Psychrobacter	2.59	6.08	4.05
Microbacterium	2.79	0.51	0.34
Agrococcus	2.74	0.00	0.00
Sphingomonas	2.51	1.52	2.20
Arthrobacter	2.42	0.00	0.00
Dokdonia	2.35	0.00	0.00
Paracoccus	2.31	2.87	4.22
Polaribacter	2.15	8.78	9.12
Roseovarius	1.94	0.00	0.00
Cobetia	1.94	5.24	6.25
Winogradskyella	1.91	1.35	2.03
Roseobacter	1.90	0.00	1.35
Cellulophaga	1.77	3.38	2.53
Mariniflexile	1.74	1.86	1.01
Pseudoalteromonas	1.71	11.82	11.99
Hoeflea	1.57	0.00	1.86
Bosea	1.30	0.68	0.84
Imperialibacter	1.23	0.00	0.00
Rhodopirellula	1.15	1.69	0.51
Dinoroseobacter	0.68	0.51	0.17
unclass. *Acidimicrobiales*	-	0.34	0.00
unclass. *Vibrionaceae*	-	11.49	10.81
unclass. *Sphingobacteriales*	-	5.24	6.93
unclass. *Micrococcaceae*	-	3.89	3.04
unclass. *Flavobacteriaceae*	-	5.41	4.05
unclass. *Rhodobacteraceae*	-	1.52	3.21
Krokinobacter	-	1.86	2.03
unclass. *Gammaproteobacteria*	-	0.34	0.34
unclass. *Actinobacteria*	-	0.34	0.00
unclass. *Proteobacteria*	-	0.17	0.00
unclass. *Alteromonadaceae*	-	0.00	0.17
unclass *Hyphomonadaceae*	-	0.34	0.00
Correlation coefficient Theory vs. Observed for genera present in the mock community		0.39	0.38