Rebuttal for "A quest for the biological sources of the ubiquitous long chain alkyl diols in the marine realm" by Sergio Balzano et al.

Comments by referee in bold; responses in normal font.

Referee 1: J. K. Volkman

Balzano et al. report an attempt to identify the sources of long-chain alkyl diols in samples of marine particulate matter in the tropical North Atlantic. The major component that they observe, in common with most marine samples, is the C30 1,15-diol. However, this is not the main chain-length in microalgae known to produce alkyl diols. For example, in marine eustigmatophytes the major diol is the C32 1,15-diol and this is accompanied by other long-chain components such as n-alkenols and unsaturated alkyl diols. The use of 18S rDNA to identify possible sources of organic matter has been successful in other studies and thus the rationale for combining genetic and biomarker data is soundly based. The fact that a clear source could not be identified is salutary and raises useful questions as to how best to combine these techniques in future studies. The paper is well written and the datasets are extensive and nicely discussed. I support publication with some changes and corrections as set out below.

We thank Dr. Volkman for his positive comments on our manuscript. Please find below detailed answers to the comments:

1) **The Introduction provides all the background information, but the structure could be improved. The first few sentences are fine, but at line 52 the text jumps to various proxies that have been developed. I think that it would be better to move the information on possible sources (line 68) here so that the reader has a clear idea of the type of distributions found and the differences between species. This should include mention of which chain-lengths are abundant and what other biomarkers might be present. This might be incorporated into Supplementary Table S1. This Table also contains a number of unpublished results, but without detail. Some of these are surprising (e.g., Heterosigma) and it is a bit disconcerting to see them referred to as known diol producers when the information has not been published. Note that Rampen et al. (2012) were not the first authors to remark that the distributions in eustigmatophytes do not match those in marine samples (see Volkman et al., 1992).**

We re-arranged the text and we are now first describing the possible sources of LCDs (lines 58-85). We do have some information about other diol producers which has indeed not been published yet. We agree that this is unfortunate, since there are no immediate plans for a separate publication of these results, but we propose to at least list them here in the supplementary Table. They were grown by the culture collections from which the algae were ordered and analysed according to the same methods as described here and we will clarify this in Table S1 as a footnote. We will cite Volkman et al. 1992 in addition to Rampen et al., 2012.

2) **The next paragraph can then introduce the proxies and add more discussion about their limitations. Like many biomarker proxies, these are empirically based from geographically limited datasets and in some cases do not have a strong mechanistic underpinning as to why they appear to correlate with oceanographic features such as temperature and upwelling. While a source of 1,14-diols is known from the diatom genus *Proboscia* which provides an explanation for why these isomers might be abundant where *Proboscia* is abundant, our lack of knowledge of the main source of the C30 1,15 diol weakens their use as a proxy. If the source can be identified then this will allow studies to underpin their use as proxy which is another justification for the type of work reported here.**

We agree and are now discussing the proxies and their limitations in lines 100-105.

3) **In the methodology it is important to explain why base and acid hydrolysis was used rather than a simple solvent extraction. If the alkyl diols were present in polar lipids, as seems likely in Nannochloropsis, then this procedure converts them to free lipids. This is relevant to later discussion of the possible effects of non-living organic matter (detritus) on the distributions. In aged samples, one might expect higher contents of free diols due to hydrolysis/degradation of polar lipids, but the method used here unfortunately does not differentiate between free and polar forms. It is well established that alkyl diols form the backbone of algaenans made by eustigmatophytes, but it is much less clear what other lipids they might occur in. Algaenans appear to be quite stable in seawater and are an unlikely source of free alkyl diols, but the possible role of other lipids is still uncertain.**

We analysed diols by base and acid hydrolysis because we initially believed that the organic matter in our suspended particulate matter was dominated by "living cells" and fresh organic matter with minor contributions from debris. Only by analysing the results we realised that the majority of diols found here were likely derived from detritus rather than living cells. Algaenans are unlikely to be a primary source of diols but preliminary degradation results (Reiche et al., Organic Geochemistry, in press, https://doi.org/10.1016/j.orggeochem.2018.08.003) suggest that substantial amounts of diols can be released from the peripheries of *Nannochloropsis* cell wall after prolonged oxygen exposure. We do not exclude the presence of compounds other than algaenans which might source the diols found here and clarified this point in line 518.

4) **The Discussion examines provides a good account of the reasons why the DNA results do not seem to match the measured abundances of the alkyl diols. I am a little concerned at the use of "LCD" as a shorthand for a variety of unrelated long-chain diol structures. I would restrict it to the C28–C32 group. It is quite likely that a number of distinct biosynthetic pathways have evolved over time in different organisms to produce compounds that are really only superficially similar in structure. To lump all these distributions together is not really appropriate. The authors make a brief mention of other compounds found in Proboscia (line 402) and use this as evidence that this genus is an unlikely source of 1,14-diols in these particular samples. This is a useful observation. I would expand the discussion here to include other biomarkers known to be present in other producers of alkyl diols such as eustigmatophytes. Assignments of possible sources are usually much more robust when multiple biomarkers are used.**

We carefully define LCDs at the start of the introduction (Line 50-51). LCD is just a chemical compound class and it would be strange to just limit it to the C28-C32 group. In the cases where we refer to this group we use a carbon chain length indication, which seems most appropriate. We analysed carefully our GC-MS results and did not detect any eustigmatophycean biomarker such as $C_{32:2}$/$C_{32:1}$ alkenols or 15-OH-$C_{30:0}$/15-OH-$C_{32:0}$ fatty acids. We incorporated this finding in the result section (lines 225-228).

5) **The Conclusion provides a nice summary of the problems of comparing DNA and biomarkers when their relative stabilities are so different. I agree that the choice of sample is very important. All samples of marine particulate matter are mixtures of living and dead material so it is important in DNA-biomarker studies to sample waters where living biomass is high (e.g. near-surface blooms). Also, if compounds exist as polar forms in living organisms then it is desirable to examine those compounds separately from hydrolysed forms in the same way that phospholipids can give information about living bacteria in a way that total fatty acids do not.**

We acknowledge that our study is limited by the fact that free and bound lipids were in this case pooled together (lines 674-677 and 696-699). Note that recent work (Reiche et al. in press) highlighted that decaying *Nannochloropsis* biomass contains high proportions of diols present as free-lipids.

6) **Minor points: I would use the common term eustigmatophyte rather than the more cumbersome eustigmatophycean, in the same way that we use diatom rather than bacillariophycean. Line 104: change to "these analyses". Line 121: no italics for "al." Line 139: It is not clear what the statement "cyanobacteria were not taken into account" means here. Were they present (even abundant), but not counted? The authors are undoubtedly aware that cyanobacteria were once proposed as a source of alkyl diols. Line 161: no spaces around the ":". Line 177: bis not Bis Line 180: 25 m not 2 5m. Line 183: SIM is usually an abbreviation for selected ion monitoring. If only these ions were run, rather than full scan, then there is a distinct possibility that other components would not be recognized. This need clarification. Line 184: the m and z in m/z should be in italics. Line 196: it is usual to use an n-dash (–) for number ranges. Line 198 and elsewhere: use a symbol prime (') not '. Line 207: dimethyl not Dimethyl Line 230: python-based Line 235: space after ")". Lines 236, 269, 521: no space before %. Line 282 and elsewhere: use correct _ symbol. Line 282: if you use the expression "between" then you cannot state a range; either state "ranged from x to y" or "in the range x–y". Line 285: salinity now has a unit (mg/kg) and psu is no longer used. Lines 294, 297: use symbol _ not x. Line 315: use station when referring to multiple stations, but Station when referring to a single numbered station. Line 393: correlated "with" rather than "to". Line 477: Cite Volkman et al., 1992 here rather than 1999. Line 531: Indent paragraph Line 542: space after comma Line 564: detritus not debris Line 677, 733, 756, 791, 800: Damsté. Line 871: subscripts for 30 and 32.**

We prefer using the term Eustigmatophyceae and the related adjective "eustigmatophycean" because most algologists consider this group as a class rather than a phylum (http://www.algaebase.org). This is because Eustigmatophyceae are considered, along with diatoms and other photosynthetic Stramenopiles, as part of the phylum Ochrophyta. The related adjective would be "eustigmatophycean".

We agree with most of the other changes suggested and we would like to clarify few further points:

Line 139: we also enumerated Cyanobacteria and, as expected, *Prochlorococcus* and *Synechococcus* were both present at densities $\approx 10^5$ cell mL^{-1}. We did not observe any correlation between these two genera and LCDs (data not shown) and we prefer to not include these data in our manuscript because this would not add useful information to the discussion. High abundances of LCDs were previously found in the Baltic Sea during a cyanobacterial bloom (Morris and Brassell, 1988) and thought to be associated with the dominant species, *Aphanizomenon flos-aquae*. However LCDs were not detected in culture material from *A. flos-aquae* (Deleeuw et al., 1992). Cyanobacteria-harbouring aquatic ferns like *Azolla* can also contain diols, which were demonstrated to be biosynthesised by the plant itself rather than the symbionts (Speelman et al., 2009). Thus, we do not believe that cyanobacteria are involved in long chain diol biosynthesis.

Line 183: all the data shown here are related to SIM chromatograms. Some of our samples (≈20) have been also analysed by full scan. We clarified this in the method section (lines 225-228).

Referee #2

Balzano et al. describe the attempt to attribute biological sources to long chain diols (LCDs), which are compounds with a great proxy potential and widespread both in the marine water column and sediments. The study relies on the analysis of suspended particulate matter in order to compare the distribution of concentration of LCDs to environment l parameters and abundance of potential LCD-producers. Unfortunately, this approach is not successful, and little information regarding LCD production can be gained. Instead, the authors provide an interesting discussion on the suitability of the combined biomarker-genetics approach. Overall, even if the results do not allow to narrow down the biological sources of LCDs, investigation is sound, and the paper is well written. A few comments are shown below.

We thank Ref. 2 for their positive comments on our manuscript. Please find below detailed answers:

Abstract: In the abstract, the authors claim that "the contributions from two taxonomic classes to which known producers are affiliated… followed a similar trend to that of the concentrations of C30 and C32 diols". This statement seems to suggest a source relationship. However, in the manuscript the authors inform that correlation is low (l. 531) and that it might be that "co-occurrences are simply driven by other environmental conditions leading to similar spatial distributions" (l.533). In my opinion both statements are not consistent and the abstract should be rephrased in order to clearly state that no informative correlation between LCD and putative LCD-producers could be established.

The reviewer is correct that in the abstract we suggest a relationship but in the manuscript we nuance this. We rephrased the abstract to say that we did find a correlation between two OTUs from Chyrsophyceae with the major diol but that this correlation is weak and might be indirect (lines 32-37).

Abstract: In the manuscript, three scenarios are discussed that explain why the correlation between LCD and potential LCD-producers is so weak: contribution of fossil LCDs, undersampling of potential LCD-producers because of their low number of rRNA gene copies per cell or LCD being produced by other species. However, in the abstract only the first hypothesis is mentioned. In my opinion, presenting all three scenarios would strengthen the manuscript.

We agree that all three scenarios should be mentioned in the abstract and this has been included in the revised version (lines 35-40).

Discussion (from l.387 on): It is argued that the C28 1,13-diol can't be correctly interpreted because of its low abundance. However, C28 1,14-diol doesn't seem to be more abundant and is discussed with a lot of detail, and concerns regarding its abundance are not expressed.

The C_{28} 1,13 diol was only detected in 19 out of 71 samples, whereas the C_{28} 1,14 diol was found, although often in low amounts, in all samples. We compared the distribution of the C_{28} 1,14 diol with that of C_{30} 1,14 diol, since both compounds can be biosynthesised by *Proboscia* spp. (Sinninghe Damsté et al., 2003). In contrast it has been suggested that the C_{28} 1,13 diol can derive from the same organisms producing the C_{30} 1,15 diol (Rampen et al., 2014). We could not compare the distribution of these two compounds because while the former was present below detection levels in 27 % of the samples, the latter was found in all the samples. We now clarify that the difficulties in the interpretation of the C_{28} 1,13 diol are due not only to its low abundance but also to the fact that it was not detected in some of our samples (lines 425-429).

Discussion (from l.563 on): Regarding the possibility of fossil LCD contributing to the signal I have a few comments/suggestions. (1) Is there any information available on the residence time of SPM in a system like the one studied? Is the claim that LCD may accumulate as SPM for years (l. 581) consistent with such residence times? (2) Bale et al. (2018) employed very similar (or actually the same?) samples from the same location to study biological sources of cyanobacterial lipids and were quite successful. However, these lipids have also been shown to persist over longer time scales (e.g. Bauersachs et al. 2010). This should be mentioned and the difference to LCD discussed (3) I would appreciate some hypothesis on LCD production, even if they are fossil to some degree. Do the authors expect seasonal production and therefore absence of producers during sampling? Export from land/freshwater systems? Production by a small population and massive accumulation? Which are the sources fueling this hypothetical fossil pool of LCDs?

1) Unfortunately, we could not find any study reporting the residence time of SPM in the Amazon Shelf or in the area or the tropical north Atlantic Ocean studied here. Most studies focus on the turnover of dissolved organic matter rather than particulate organic matter.
2) Yes, Bale et al. (2018) analysed samples from the same (HCC) oceanographic cruise. The authors analysed intact polar lipids, which are more suitable indicators of "fresh" organic matter. We cited the recent paper of Bale et al. in lines 668-677.
3) Most producers of the LCDs measured here were indeed unlikely to be present in seawater during sampling. Whether LCD-producers occurred during another period of the year, or in other locations, is unfortunately unknown, also as we do not know who the main producers are. Export from freshwater is unlikely because only 6 stations (7-13) were slightly affected by Amazon River as shown in the salinity profile (Fig. 1). In general, the Amazon River input is low for the period of the year in which sampling took place (Molleri et al., 2010). We included this statement in lines 464-468.

Minor comments:
-l.117 (also legend for figure 1), what does HCC stand for? That is the cruise name, it stands for HeteroCystous Cyanobacteria (line 155), which was the original focus of that cruise.

-when expressing ratios of, for example, solvents (e.g. l.159 "HCl: MeOH (1:1") empty spaces before and/or after the colon are not employed consistently. Please check throughout the text. We are now not using empty spaces after the colon when reporting ratios.

-l. 184: as far as I know, it is recommended to write "m/z" (mass to charge ratio) in Italics. Done

-l**. 296. Please use either "Station" or "Stn." consistently**. Done

-Figure 1: do de dots represent sampling depths? Please explain in legend. Yes they do, the legend has been clarified (lines 1026-1027).

-Figure 1: Bale et al. (2018) used chl-a obtained by fluorescence instead of the extraction-based approach used here, and those data seem to have a better coverage/ resolution. Could you please explain why you are not using them? We preferred using Chl-*a* data based on methanol extraction and HPLC analyses as we could report these data as pg L^{-1}.

-Should Figure 4 maybe also be in colour (like Fig. 1 and 2)? Figs. 1-2 report different parameters which vary along a continuum (temperature, salinity, concentrations) whereas Fig. 4 is reporting the number of reads associated with specific taxa. We did not detect these reads in many samples, which appear in white and we do not see a specific need for using several colours.

-Consider adding P-values to figure 5. We added p-values on each plot of Fig. 5.

Referee #3

In their manuscript "A quest for the biological sources of the ubiquitous long chain alkyl diols in the marine realm", S. Balzano and co-authors present a detailed lipid-DNA comparison along a transect in the tropical North Atlantic for long-chain diols and their producers.

Long-chain diols (LCD) have been of considerable interest to the community for a few years now, and show some potential as proxies for riverine input, upwelling, or potentially temperature. As their sources have not yet been clearly identified, it is timely to use an approach to combine molecular biology and DNA, which the authors employed in this study. Balzano and co-authors, using an in situ filtering approach, found diol concentrations as expected at this site, but were not able to detect the DNA of enough diol producers to account for the amounts of diols detected.

The research presented is thorough, and the manuscript is clear. I have a few questions and comments:

We thank Ref. 3 for their positive comments on our manuscript. Please find below detailed answers:

The title is engaging, but sounds more like a general review of the topic, and does not reflect the content at all. Rephrase this to clearly indicate the study area and the results.

We rephrased the title which is now "A quest for the biological sources of long chain alkyl diols in the western tropical North Atlantic Ocean".

Why was this specific study area selected, what makes it useful for the research question?

We selected samples from the HCC cruise for this study because the stations sampled include two groups of off-shore stations (1-6 and 15-23) separated by some stations slightly affected by the Amazon River (7-13) as shown in Fig. 1. The original target of the cruise was to sample for heterocystous cyanobacteria which were known to occur in this area (see Bale et al., 2018).

There is a mismatch between the DNA and the diol concentrations. Could this be because of the size fraction sampled (0.7 _m)? This is addressed (I think) by the comparison of the cell counts and the discussion in L599-606, but should be made clearer.

We sampled all the eukaryotic plankton (> 0.7 µm) and used the same filter to analyse both LCDs and microbes. We do not see any bias in this approach and we believe that there are three main reasons for the mismatch between microbial community and diol distribution: (1) primer mismatch leading to real LCD-producers being undetected, (2) fossil nature LCDs and (3) undersampling of potential LCD-producers because of their very low number of rRNA gene copies. We have clarified this in the discussion (lines 506-521 and 612-632).

The supplementary data is great and detailed, but the diol concentrations should be added in a table as well. Have the sequences been deposited in GenBank

We added the raw diol concentrations in Table S2 as well as the Genbank accession number (SUB4388921) for the sequences

The references are inconsistent, some contain a doi, some don't, some include the doi as a link, the citation for ODV is not correct.

We checked all the references, added the doi where available and made sure they are all in the correct format for Biogeosciences

L182-183: The temperature regime is a very minor detail to add, so for the sake of the reader who wants to reproduce this, I would add this to the section.

The temperatures of the sampling points analysed here are also shown in Fig. 1A.

L277-297: Is this new data or has this been published by Bale et al. 2017?

Some of these data are also published in Bale et al. (2018), since both manuscripts are referring to samples collected during the same cruise. We clarified in the manuscript which data are from Bale et al. (2018) and which once are published here for the first time (lines 159-160, see also legend of Fig. 1).

Table 2: Is that % abundance or actual concentrations?

Table 1 refers to the number of rRNA gene reads from different taxa as well as their percentage contribution to the total rRNA gene reads. We clarified this in Table 1. Table 2 and Table 3 show instead Spearman correlation values.

Figure 1&2: Considering the two-dimensional nature of the transect, a supplementary online 3D plot could be useful.

We believe that Fig. 1 and Fig. 2 show sufficient information on the physical parameters and the LCD concentrations on the samples analyses.

References

Bale, N. J., Villareal, T. A., Hopmans, E. C., Brussaard, C. P. D., Besseling, M., Dorhout, D., Sinninghe Damsté, J. S., and Schouten, S.: C5 glycolipids of heterocystous cyanobacteria track symbiont abundance in the diatom *Hemiaulus hauckii* across the tropical north Atlantic, Biogeosciences, 15, 1229–1241, https://doi.org/10.5194/bg-2017-300, 2018.

Deleeuw, J. W., Rijpstra, W. I. C., and Mur, L. R.: The absence of long-chain alkyl diols and alkyl keto-1-ols in cultures of the cyanobacterium *Aphanizomenon-flos-aquae*, Org. Geochem., 18, 575-578, 10.1016/0146-6380(92)90120-m, 1992.

Molleri, G. S. F., Novo, E., and Kampel, M.: Space-time variability of the Amazon River plume based on satellite ocean color, Cont. Shelf Res., 30, 342-352, https://doi.org/10.1016/j.csr.2009.11.015, 2010.

Morris, R. J., and Brassell, S. C.: Long chain alkanediols. Biological markers for cyanobacterial contributions to sediments. , Lipids, 23, 256-258, 10.1007/bf02535468, 1988.

Rampen, S. W., Datema, M., Rodrigo-Gamiz, M., Schouten, S., Reichart, G.-J., and Sinninghe Damsté, J. S.: Sources and proxy potential of long chain alkyl diols in lacustrine environments, Geochim. Cosmochim. Ac., 144, 59-71, https://doi.org/10.1016/j.gca.2014.08.033, 2014.

Sinninghe Damsté, J. S., Rampen, S., Irene, W., Rupstra, C., Abbas, B., Muyzer, G., and Schouten, S.: A diatomaceous origin for long-chain diols and mid-chain hydroxy methyl alkanoates widely occurring in Quaternary marine sediments: Indicators for high-nutrient conditions, Geochim. Cosmochim. Ac., 67, 1339-1348, https://doi.org/10.1016/s0016-7037(02)01225-5, 2003.

Speelman, E. N., Reichart, G. J., de Leeuw, J. W., Rijpstra, W. I. C., and Sinninghe Damste, J. S.: Biomarker lipids of the freshwater fern Azolla and its fossil counterpart from the Eocene Arctic Ocean, Org. Geochem., 40, 628-637, https://doi.org/10.1016/j.orggeochem.2009.02.001, 2009.

A quest for the biological sources of long chain alkyl diols in the western tropical North Atlantic Ocean.

Sergio Balzano[a*], Julie Lattaud[a], Laura Villanueva[a], Sebastiaan Rampen[a], Corina P.D. Brussaard[a], Judith van Bleijswijk[a], Nicole Bale[a], Jaap S. Sinninghe Damsté[a,b], Stefan Schouten[a,b]

[a] Department of Marine Microbiology and Biogeochemistry (MMB) and Utrecht University, NIOZ Royal Netherlands Institute for Sea Research, PO Box 59, 1790 AB Den Burg, The Netherlands

[b] Department of Earth Sciences, Faculty of Geosciences, Utrecht University, Budapestlaan 4, 3584 CD Utrecht, The Netherlands

* Correspondence to Sergio Balzano. E-mail address: Sergio.balzano@nioz.nl (S. Balzano).

Abstract

Long chain alkyl diols (LCDs) are widespread in the marine water column and sediments but their biological sources are mostly unknown. Here we combine lipid analyses with 18S rRNA gene amplicon sequencing on suspended particulate matter (SPM) collected in the photic zone of the western tropical North Atlantic Ocean at 24 stations to infer relationships between LCDs and potential LCD-producers. The C_{30} 1,15-diol was detected in all SPM samples and accounted for >95% of the total LCDs, while minor proportions of C_{28} and C_{30} 1,13-diols, C_{28} and C_{30} 1,14-diols as well as C_{32} 1,15-diol were found. The concentration of the C_{30} and C_{32} diols was higher in the mixed layer of the water column compared to the deep chlorophyll maximum (DCM), whereas concentrations of C_{28} diols were comparable. Sequencing analyses revealed extremely low contributions (≈ 0.1% of the 18S rRNA gene reads) of known LCD-producers but the contributions from two taxonomic classes to which known producers are affiliated, i.e. Dictyochophyceae and Chrysophyceae, followed a trend similar to that of the concentrations of C_{30} and C_{32} diols. Statistical analyses indicated that the abundance of 4 operational taxonomic units (OTUs) of the Chrysophyceae and Dictyochophyceae, along with 23 OTUs falling in other phylogenetic groups, were weakly ($r \leq 0.6$) but significantly (p-value < 0.01) correlated with C_{30} diol concentrations. It is not clear whether some of these OTUs might indeed correspond to $C_{28\text{-}32}$ diol-producers or whether these correlations are just indirect and the occurrence of C_{30} diols and specific OTUs in the same samples might be driven by other environmental conditions. Moreover, primer mismatches were unlikely but cannot be excluded and the variable number of rRNA gene copies within eukaryotes, might have affected the analyses leading to LCD-producers being undetected or under sampled. Furthermore, based on the average LCD-content measured in cultivated LCD-producing algae, the detected concentrations of LCDs in SPM are too high to be explained by the abundances of the suspected LCD-producing OTUs. This is likely explained by the slower degradation of LCDs compared to DNA in the oxic water column and suggests that some of the LCDs found here were likely to be associated to suspended debris, while the DNA from the related LCD-producers had been already fully degraded. This suggests that care should be taken in constraining biological sources of relatively stable biomarker lipids by quantitative comparisons of DNA and lipid abundances.

Keywords: long chain alkyl diols, HCC, tropical North Atlantic, 18S rRNA gene amplicon sequencing, Eustigmatophyceae, Chrysophyceae, Dictyochophyceae

1. INTRODUCTION

Long chain alkyl diols (LCDs) are lipids that consist of a linear alkyl chain with 22–38 carbons, hydroxylated at both the terminal carbon atom and at an intermediate position, and usually saturated or monounsaturated. LCDs were identified for the first time in Black Sea sediments (de Leeuw et al., 1981) and have subsequently been found with widespread occurrence in both suspended particulate matter (SPM) and sediments from both coastal and off-shore sites throughout the World Ocean (Jiang et al., 1994; Versteegh et al., 1997; Rampen et al., 2014b). LCDs can be preserved in marine sediments for long periods of time and their distribution can reflect the environmental conditions at the time they were produced.

The most abundant LCDs in seawater are the saturated C_{28} and C_{30} 1,13-diols, C_{28} and C_{30} 1,14-diols, and C_{30} and C_{32} 1,15-diols (Rampen et al., 2014b), which are all likely produced by phytoplankton. However, the marine biological sources of LCDs are still not fully clear because, in contrast with the widespread occurrence of LCDs in the sediment, few marine taxa have been shown to contain these lipids. Eustigmatophyceae contain C_{30} 1,13-, C_{30} 1,15-, and C_{32} 1,15-diols (Volkman et al., 1992; Rampen et al., 2014a) but they comprise mostly freshwater species and only a few rare marine representatives from the genus *Nannochloropsis* are known (Andersen et al., 1998; Fawley and Fawley, 2007). Furthermore, the distribution of LCDs in the marine environment does not match that of LCDs of marine Eustigmatophyceae (Volkman et al., 1992; Rampen et al., 2012). Species of the diatom genera *Proboscia* and the dictyocophycean *Apedinella radians* contain C_{28-32} 1,14-diols (Sinninghe Damsté et al., 2003; Rampen et al., 2009; Rampen et al., 2011), with the former accounting for significant proportions of marine biomass mostly in upwelling regions (Moita et al., 2003; Lassiter et al., 2006), whereas the latter has been occasionally observed in estuarine environments (Seoane et al., 2005; Bergesch et al., 2008). Few other marine species from classes genetically related to diatoms and Eustigmatophyceae have been recently shown to produce LCDs (Table S1). All the known LCD-producing phytoplankters belong to the eukaryotic supergroup Heterokontophyta, a division which includes, among others, diatoms and brown seaweeds. The widespread occurrence of LCDs in the marine environment, ~~in spite of~~despite the restricted abundance and distribution of marine ~~LCD-producerLCD~~LCD-producers~~s~~, suggests that these compounds may be produced by unknown phytoplankton species. In addition LCD in the marine environment might also derive from vegetal debris of terrestrial or riverine origin. For example, C_{30-36} diols functionalised at the 1- and the ω18 or ω20 positions have previously been reported to occur in ferns (Jetter and Riederer, 1999;

Speelman et al., 2009; Mao et al., 2017) and suggested to be part of the leaf cuticular waxes. Similarly, C_{26-32} diols have been occasionally detected in plants (Buschhaus et al., 2013). This suggests that vegetal debris may in principle also source LCDs in seawater.

Several indices, based on ratios between the different diols, have been proposed for the reconstruction of past environmental conditions. The Diol Index, reflecting the proportion of C_{28} and C_{30} 1,14-diols over the sum of C_{28} and C_{30} 1,14-diols and C_{30} 1,15-diol, has been proposed to track ancient upwelling conditions since the 1,14-diols are believed to be mostly produced by upwelling diatoms of the genus *Proboscia* (Rampen et al., 2008). Another index, the long chain diol index (LDI), which is based on the proportion of the C_{30} 1,15-diol over the C_{28} and C_{30} 1,13-diols, shows a strong correlation with sea surface temperature (SST) and is used to determine past SST (Rampen et al., 2012; Plancq et al., 2014; Rodrigo-Gámiz et al., 2015). In addition, since the C_{32} 1,15-diol is the major component of the LCDs of freshwater Eustigmatophyceae (Volkman et al., 1992; Rampen et al., 2014a), the fractional abundance of C_{32} 1,15-diol has been suggested to be a marker of riverine input in seawater (de Bar et al., 2016; Lattaud et al., 2017a; Lattaud et al., 2017b). Other markers for riverine inputs in seawater are the C_{30-36} 1,ω20-diols which are produced by the freshwater fern *Azolla* (Speelman et al., 2009; Mao et al., 2017). However, application of these proxies in the marine realm remains uncertain. For example the growth of *Proboscia* spp. is typically promoted under low concentrations of dissolved silica, whereas other diatoms dominate upwelling area under higher silica concentrations (Koning et al., 2001), making the Diol Index ineffective in predicting upwelling conditions when communities are dominated by other diatoms. In addition, the sources of the major marine C_{30} 1,15-diol are unknown, complicating the application of the LDI as a proxy.~~.~~

A way of assessing the sources of biomarker lipids is to compare the abundance of lipids in environmental samples with the composition of the microbial community, as determined by genetic methods. For example, Villanueva et al. (2014) analysed both LCDs and eustigmatophycean 18S rRNA gene sequences in a tropical freshwater lake and found five clades of uncultured Eustigmatophyceae in the top 25 m of the water column of the lake, where LCDs were also abundant. Abundance determination by quantitative polymerase chain reaction (qPCR) highlighted that the number of eustigmatophycean 18S rRNA gene copies peaked at the same depth as the LCDs, suggesting that Eustigmatophyceae are a primary source for LCDs in freshwater (Villanueva et al., 2014). However, one of the limitations of this approach is that it relies on specific eustigmatophycean primers designed based on the sequences available in the genetic databases, which could be biased and not target all the

existing LCD biological sources. To compensate for this limitation high throughput amplicon sequencing of the 18S rRNA gene allows the exploration of the total marine microbial communities in great detail (Stoeck et al., 2009; Logares et al., 2012; Christaki et al., 2014; Balzano et al., 2015; de Vargas et al., 2015; Massana et al., 2015). The combination of these analyses with lipid composition may potentially assist in identifying the main LCD producers in marine settings.

In the present study, we quantitatively analysed the composition and abundance of LCDs in suspended particulate matter (SPM) collected along the tropical North Atlantic (Fig. 1A) at different depths in the photic zone (surface, deep chlorophyll maximum (DCM) and bottom of the wind mixed layer (BWML); see also Bale et al., 2018). The 18S rRNA gene abundance and composition of the SPM was also analysed by quantitative PCR (qPCR) and high throughput amplicon sequencing to infer the taxonomic composition and to compare the abundance of the different taxa with that of the LCDs, in order to identify the potential marine biological sources of LCDs.

2. MATERIAL AND METHODS

2.1 Cruise transect, ancillary data, and SPM collection

Samples were taken during the Heterocystous Cyanobacteria Cruise (HCC) (64PE393), which took place from 24th August to 21st September 2014 along a transect on the tropical North Atlantic Ocean (see Bale et al. (2018) for details). The transect was from Mindelo (Cape Verde) to a location about 500 km from the Amazon River mouth and then westwards along the coast towards Barbados (Fig. 1A). Temperature, salinity and nutrient data have previously been reported in Bale et al. (2018).

Seawater was collected from two or three depths at each station to measure the concentration of chlorophyll *a* (Chl-*a*) and the abundances of photosynthetic pico and nanoeukaryotes. Seawater was collected during the up cast using Niskin bottles mounted on a CTD frame. The sampling depths were determined based on the evaluation of the vertical profiles of temperature, salinity, and chlorophyll fluorescence after the down cast of the CTD deployment. The depth of the BWML and the DCM were determined based on the lowest position of the mixed layer and the depth at which the highest values of chlorophyll fluorescence were observed. For Chl-*a* determination seawater was collected from the Niskin bottles and filtered through 0.7 µm pore-size glass-fiber (Whatman GF/F) filters, followed by frozen storage. Chl-*a* was extracted with methanol buffered with 0.5 M ammonium acetate, homogenized for 15 s and analysed by high performance liquid chromatography.

Photosynthetic pico- and nanoeukaryotes were enumerated by flow cytometry according to the protocol of Marie et al. (2005). In short, 1 mL samples were counted fresh using a Becton-Dickinson FACSCalibur (Erembodegem, Belgium) flow cytometer equipped with an air-cooled Argon laser (488 nm, 15 mW). Phytoplankton were discriminated based on their chlorophyll autofluorescence and scatter signature. Cyanobacteria, i.e. *Synechococcus* and *Prochlorococcus*, were not included in the current study. Size fractionation was performed by gravity filtration with >3 µm average cell diameter phytoplankton groups classified as nanoeukaryotic and those <3 µm average cell diameter as picoeukaryotic phytoplankton.

Three McLane *in situ* pumps (McLane Laboratories Inc., Falmouth) were used to collect SPM from the water column for the analysis of both lipids and microbial communities. As with the collection of seawater with Niskin bottles for Chl-*a* and flow cytometry analyses, the *in situ* pumps were deployed at the surface (3 - 5 m depth), the BWML and the DCM (Table S2). Between 100 and 400 L of seawater was pumped and the SPM was collected on pre-

combusted 0.7 µm GF/F filters (Pall Corporation, Washington) and immediately frozen at -80°C. For the determination of the organic carbon concentrations, SPM was freeze dried and analysis was carried out using a Flash 2000 series Elemental Analyzer (Thermo Scientific) equipped with a thermal conductivity detector.

2.2 Lipid extraction and analyses of LCDs

Lipids were extracted from the GF/F filters as described previously (Lattaud et al., 2017b). Briefly, ¼ of the filters were dried using a LyoQuest (Telstart, Life Sciences) freeze-dryer and lipids were extracted using base and acid hydrolysis. The base hydrolysis was achieved with 12 mL of a 1 M KOH in methanol solution by refluxing for 1 h. Subsequently, the pH was adjusted to 4 with 2 M HCl:CH_3OH (1:1, v/v) and the extract was transferred into a separatory funnel. The residues were further extracted once with CH_3OH:H_2O (1:1, v/v), twice with CH_3OH, and three times with dichloromethane (DCM). The extracts were combined in the separatory funnel and bidistilled water (6 mL) was added. The combined solutions were mixed, shaken and separated into a CH_3OH:H_2O and a DCM phase, the DCM phase was removed and collected in a centrifuge tube. The aqueous layer was re-extracted twice with 3 mL DCM. The pooled DCM layers were dried over a sodium sulfate column and the DCM was evaporated under a stream of nitrogen. The extract was then acid hydrolyzed with 2 mL of 1.5 M HCl in CH_3OH solution under reflux for 2 h. The pH was adjusted to 4 by adding 2 M KOH:CH_3OH. 2 mL of DCM and 2 mL of bidistilled water were added to the hydrolyzed extract, mixed and shaken and, after phase separation, the DCM layer was transferred into another centrifuge tube. The remaining aqueous layer was washed twice with 2 mL DCM. The combined DCM layers were dried over a sodium sulfate column, the DCM was evaporated under a stream of nitrogen and a C_{22} 5,17-diol was added to the extract as internal standard. The extract was separated on an activated aluminium oxide column into three fractions using the following solvents: hexane:DCM (9:1, v/v), hexane:DCM (1:1, v/v) and DCM:CH_3OH (1:1, v/v). The latter (polar) fraction containing the diols was dried under a gentle nitrogen stream. Diols were derivatized by silylating an aliquot of the polar fraction with 10 µL N,O-bis(trimethylsilyl) trifluoroacetamide (BSTFA) and 10 µL pyridine, heating for 30 min at 60 ºC and adding 30 µL of ethyl acetate. The analysis of diols was performed by gas chromatography-mass spectrometry (GC-MS) using an Agilent 7990B GC gas chromatograph, equipped with a fused silica capillary column (2-5 m x 320 µm) coated with CP Sil-5 (film thickness 0.12 µm), coupled to an Agilent 5977A MSD mass spectrometer. The temperature regime for the oven was the same as that used by Lattaud et al. (2017b): held

at 70 °C for 1 min, increased to 130 °C at a rate of~~for~~ 20 °C min^{-1}, increased to 320 °C at a rate of 4 °C min^{-1}, held at 320 °C for 25 min. The flow was held constant at 2 mL min^{-1}. The MS source temperature was held at 250 °C and the MS quadrupole at 150 °C. The diols were identified and quantified via Single Ion Monitoring (SIM) of the *m/z*= 299.3 (C_{28} 1,14-diol), 313.3 (C_{28} 1,13-diol, C_{30} 1,15-diol), 327.3 (C_{30} 1,14-diol) and 341.3 (C_{30} 1,13-diol, C_{32} 1,15-diol) ions (Versteegh et al., 1997; Rampen et al., 2012). Surface samples, which contained the highest concentrations of LCDs, were also analysed by full scan to evaluate the presence of other eustigmatophycean biomarkers such as long chain alkenols and long chain hydroxy fatty acids. Absolute concentrations were calculated using the peak area of the internal standard as a reference.

2.3 DNA extraction, PCR, qPCR, and 18S rRNA gene sequencing

On ice a small portion of the GF/F filters, corresponding to 1/16 of their initial size, hence containing SPM from ca. 25 L of seawater, was cut into small pieces using sterile scissors and tweezers. Filter pieces were then transferred into 2 mL microtubes and the DNA was extracted using a MOBIO powersoil DNA isolation kit (Qiagen) following manufacturer instructions. We amplified the hypervariable V4 region of the 18S rRNA which is considered the best genetic marker for the identification of microbial eukaryotes (Logares et al., 2012; Massana et al., 2015). The V4 is located in a central region (565-584 bp to 964-981 bp for *Saccharomyces cerevisiae*) of the 18S rRNA and it was amplified from the genomic DNA by PCR using the universal eukaryotic primers TAReuk454FWD1 (5´-CCAGCASCYGCGGTAATTCC-3´) and TAReuk454REV3 (5´-ACTTTCGTTCTTGATYRA-3´) (Stoeck et al., 2010). Primers were modified for multiplex sequencing on a Roche 454 GS FLX system: a 454-adapter A (CCATCTCATCCCTGCGTGTCTCCGACTCAG), a key (TCAG), and a 10 bp sample-specific Multiple Identifier (MID, Table S3) were bound to the 5´ end of the forward primer, whereas a 454-adapter 2 (CCTATCCCCTGTGTGCCTTGGCAGTCTCAG) and a unique MID (CGTGTCA) were bound to the 5´ end of the reverse primer for all the samples. The PCR mixture included 25 µL Phusion Flash High-Fidelity PCR Master Mix (ThermoFisher Scientific) 19.1 µL deionised water, 1.5 µL dimethyl sulfoxide, 1.7 µL from each primer and 25 ng genomic DNA and the V4 region was amplified using the same thermal cycling as described by Logares et al. (2012). Amplicons were visualised on a 1% agarose gel and V4 bands were excised and subsequently purified using a QIAquick Gel Extraction Kit (Qiagen) and DNA concentration was measured by Qubit Fluorometric Quantitation (ThermoFisher

Scientific). For each sequencing run, 20 samples were pooled in equimolar amount and sequenced using a 454 GS-FLX Plus (Macrogen Korea). Some samples yielded a low number of reads and were re-sequenced; overall 77 samples were sequenced in 5 sequencing runs.

To determine the concentration of total 18S rRNA genes within the seawater sampled we carried out qPCR using the same primers and the same cycling conditions as described above. qPCR analysis was performed on a Biorad CFX96TM Real-Time System/C1000 Thermal cycler equipped with CFX Manager^{TM} Software. Abundance of 18S rRNA gene sequences was determined with the same primer pair (TAReuk454FWD1/ TAReuk454REV3) used for the 18S rRNA gene diversity analysis. Each reaction contained 12.5 µL MasterMix phusion, 8.25 µL deionised nuclease-free water, 0.75 µL DMSO, 1 µL from each primer and 0.5 µL Sybr green and 1 µL of DNA template. Reactions were performed in iCycler iQTM 96-well plates (Bio-Rad). A mixture of V4 18S rRNA gene amplicons obtained as described above was used to prepare standard solutions. All qPCR reactions were performed in triplicate with standard curves from 6.4x10^3 to 6.4x10^9 V4 molecules per microliter. Specificity of the qPCR was verified with melting curve analyses (50 °C to 95 °C).

2.4 Bioinformatic analyses

Bioinformatic analyses were carried out using the python-based bioinformatic pipeline quantitative insight in microbial ecology (QIIME) (Caporaso et al., 2010). Overall, we obtained 372 107 raw sequences; reads with a length comprised between 250 and 500 bp, less than 8 homopolymers, and a phred quality ≥ 25 over 50 bp sliding windows were kept for downstream analyses. Chimeric sequences were then identified by comparison with the Protist Ribosomal Database 2 (PR2) (Guillou et al., 2013) using the Uchime algorithm (Edgar et al., 2011) and removed from the dataset along with singletons (i.e. reads not sharing 100% identity with at least another read).

A total of 238 564 reads remaining after quality filtering were clustered into 2457 Operational Taxonomic Units (OTUs) based on 95% sequence identity using Uclust (Edgar, 2010). Samples containing less than 1000 sequencing reads were removed from the dataset. The taxonomic affiliation of the OTUs was then inferred by comparison with the PR2 (Guillou et al., 2013) using BLAST (Altschul et al., 1990) within the QIIME pipeline. Reads from metazoa and multicellular fungi were removed from the dataset which finally contained 1871 OTUs and 184 279 reads. A representative set of sequences from the OTUs used here has been submitted to the GenBank (SUB4388921).The abundance of the different taxa in each sample were estimated by multiplying the percentage of reads with the concentration of

V4 copies measured by qPCR. Taxa containing $C_{28\text{-}32}$ diol-producers were extracted from the dataset and plotted using Ocean Data View (ODV) (Schlitzer, 2002).

2.5 Statistical analyses

Linear regression analyses between the concentrations of the different LCDs were performed to assess whether some of the LCDs were likely to derive from a common source. To investigate relationships between LCDs and environmental conditions we calculated the Spearman rank correlation coefficient (r) using the R package vegan (Dixon, 2003). The environmental data used were temperature, salinity, TOC, nutrients (nitrate, nitrite, ammonium, phosphate, and silica), as well the concentration of Chl-*a* and the abundance of photosynthetic pico and nanoeukaryotes. Samples containing missing data and outliers were removed from the dataset before the calculations. Both correlation coefficients and p-values were calculated and the latter were corrected for False Discovery Rates (Benjamini and Hochberg, 1995). Correlations were considered significant for p-values <0.01.

To investigate the relationships between lipids and microbial taxa we also calculated the Spearman's rank correlation coefficient between the LCD concentrations and the abundance of the different taxa at both OTU and class levels. To this end, taxonomic data were normalized based on the number of V4 copies in the different samples measured by qPCR.

Comparisons at class level provide the advantage of pooling distribution data from several closely-related OTUs, thus reducing the number of zeros (samples where a given OTU is absent), which complicates statistical analyses of biological distributions (Legendre and Gallagher, 2001). However pooling OTUs at higher taxonomic levels likely leads to combining of species able and unable to produce LCDs falling into the same taxonomic level. We thus removed OTUs that were observed in fewer than 19 samples (25%) and compared the resulting OTU table with the LCD concentrations. These analyses were performed using the qiime script *observation_metadata_correlation.py* (Caporaso et al., 2010) and the p-values were corrected for false discovery rates (Benjamini and Hochberg, 1995).

3. RESULTS

3.1 Ancillary data

The HCC cruise sailed across tropical Atlantic waters (Fig. 1A) in late summer and was targeted at SPM from the photic zone collected at the surface, the BWML and the DCM. The extent of the photic zone as well as the depths of both BWML and DCM at each station were assessed based on the vertical profiles of temperature, salinity and chlorophyll fluorescence. The temperature of photic zone waters ranged from 15 to 29 °C (Fig. 1B, Table S2), the BWML depth was comprised between 9 and 40 m, whereas the depth of the DCM ranged from 45 to 105 m. Temperatures varied at the DCM increasing westwards, whereas they were relatively constant at surface and BWML. Salinity varied between 29 and 36.5 g kg^{-1} (Fig. 1C, Table S2) at the surface, whereas it was fairly constant in the DCM (36 to 37). The concentration of Chl-*a* varied from 34 to 470 ng L^{-1} (Fig. 1D, Table S2), with the lowest values measured at the surface of the easternmost (1 to 6) and westernmost (21 to 23) stations and the relatively higher concentrations in surface waters of the shallowest stations (11 to 13) located above the continental shelf and about 500 km off the Amazon River mouth (Fig. 1A, Table S2). The POC concentration ranged from 0.6 to 13 mg L^{-1} and also peaked at surface for the shallowest stations (Fig. 1E, Table S2).

Photosynthetic picoeukaryotes, quantified by flow cytometry, were more abundant at the DCM compared to surface and BWML (Fig. 1F). Their abundance peaked at the DCM of Stations 1 and 2 ($>1.5 \cdot 10^7$ cell L^{-1}), whereas for surface waters the highest values were measured at Stations 11 to 13. In contrast, photosynthetic nanoeukaryotes did not vary substantially through the water column and their abundance peaked at the surface of Station 17 reaching a density of $1.4 \cdot 10^5$ cell L^{-1} (Fig. 1G).

3.2 Long chain alkyl diols

Six LCDs were detected, the C_{28} and C_{30} 1,13-diols, C_{28} and C_{30} 1,14-diols and the C_{30} and C_{32} 1,15-diols (Fig. 2, Table S2). The C_{30} 1,15-diol dominated all samples, accounting for >95% of the total LCDs, and its concentration ranged from 100 to 1600 pg L^{-1}. The concentration of the C_{28} 1,13-diol ranged from 0 (i.e. undetectable) to 55 pg L^{-1}, whereas the highest concentration measured for the C_{28} 1,14-diol was 64 pg L^{-1}. The other minor diols were usually more abundant than the C_{28} diols, reaching concentrations up to 190 pg L^{-1} for the C_{30} 1,13-diol, 240 pg L^{-1} for the C_{30} 1,14-diol, and 480 pg L^{-1} for the C_{32} 1,15-diol (Fig. 2).

The concentration of the C_{28} 1,13-diol peaked in the surface waters of Station 10, but it was below the detection limit in 19 samples from different depths and stations (Fig. 2A). The C_{28} 1,14-diol reached its highest concentrations at the DCM of Station 12 (64 pg L^{-1}) and at the surface of Station 13 (45 pg L^{-1}) and tended to be more abundant in the waters of the eastern stations (Fig. 2B). The concentrations of both C_{28} 1,13- and C_{28} 1,14-diols did not vary significantly with depth (t-test, p-value >0.1), while those of the C_{30} 1,13-, C_{30} 1,14-, and C_{30} 1,15-diols were higher in the mixed layer (surface and BWML) compared to the DCM (p-value <0.01).

The concentration of the C_{30} 1,13-diol peaked at the surface of Stations 10 and 14 (Fig. 2C), while that of the C_{30} 1,14-diol reached its maximum at the BWML of Stations 7 and 8 (Fig. 2D). The highest concentration of the C_{30} 1,15-diol was measured at the surface of Station 17 (16 ng L^{-1}, Fig. 2E). The concentration of the C_{32} 1,15-diol peaked in the surface waters of Stations 10 and 14 and at the DCM of Station 7 (Fig. 2F) and its concentration did not vary significantly with depth. The concentrations of both the C_{30} and C_{32} diols peaked in the mixed layer of Stations 7-10 and 14-17, which are located in close proximity to the Amazon Shelf (Figs. 2C-F).

3.3 Eukaryotic 18S rRNA gene diversity analysis

Sequencing of the hypervariable V4 region of the 18S rRNA gene of 68 SPM samples resulted in 238 564 reads with an average of 4 987 reads per sample (Table S2). Reads were clustered based on 95% sequence identity and, after removal of reads of Metazoa and multicellular fungi, we obtained 1871 operational taxonomic units (OTUs). Rarefaction analyses indicate that >90% of the genetic diversity was captured (Fig. S1), suggesting that no sample was under sequenced. Most (>90%) reads sequenced here were assigned to Dinophyceae, Syndiniales, Metazoa, Haptophyta, and Radiolaria (Fig. 3). Samples were grouped according to the depth layer (surface, BWML, and DCM) and analysis of similarity (anosim) revealed that the average variance between samples from different groups was higher than the average variance between samples from the same group (p-value ≈ 0.001), indicating that the eukaryotic community was mostly influenced by the water depth rather than the geographic location. The proportion of reads from Dinophyceae, Syndiniales, and Haptophyta was slightly higher in the mixed layer compared to the DCM, whereas Radiolaria and Pelagophyceae tended to be slightly more abundant in deeper waters (Fig. 3). All samples except surface waters from Station 12, the BWML from Station 11 and the DCM from Station 22 exhibited high contributions (>50%) from Dinophyceae and Syndiniales (Fig. S2).

Radiolaria dominated the DCM at Station 22, diatoms were relatively abundant (≈ 10-20%) at the surface of Stations 12-14 and the BWML of Station 12 while the contribution of diatom reads was <5% for all the other samples.

18S rRNA gene reads of only four taxa containing known LCD-producers were detected within our dataset: *Proboscia* spp., Florenciellales, *Heterosigma* spp., and Eustigmatophyceae (Table 1). In 33 out of 68 SPM samples we did not detect any 18S rRNA gene read from known LCD-producers, whereas reads from these taxa accounted for <0.1% of the total 18S rRNA reads in 24 samples, 0.1 to 0.5% in 8 samples, 0.5 to 1% in 2 samples and 1.5% in one sample (Station 20, BWML). The 18S rRNA gene reads from putative LCD-producers were mostly recovered from the mixed layer (Table 1). Florenciellales was the most abundant taxon among the known LCD-producers since it exhibited the highest number of reads (99) and was present in 28 out of 68 samples. The other taxa of putative LCD-producers were detected only in 8 (Eustigmatophyceae) or 2 (*Proboscia* sp. and *Heterosigma akashiwo*) samples (Table 1) accounting from 3 (*Proboscia*) to 45 (Eustigmatophyceae) reads. Eustigmatophyceae (mostly affiliated to *Nannochloropsis oculata*) were found at surface for the Stations 11, 12, and 13, as well as at the DCM of Station 20 (Fig. 4A).

Since species genetically related to cultivated microalgae known to produce LCD may also contain LCDs, we expanded our community composition analyses to groups at a higher taxonomic level and focused on those classes or divisions that contain LCD-producers (Table S1). Specifically we investigated the distribution of Eustigmatophyceae, since they are the most well-known class of LCD-producers, Pelagophyceae and Chrysophyceae, which include the LCD-producers *Sarcinochrysis marina* and *Chrysosphaera parvula*, respectively (Table S1), Dictyochophyceae, which includes *Apedinella radians* (Rampen et al., 2011), and Raphidophyceae, which include two LCD-producers, *H. akashiwo* and *Haramonas dimorpha*. We did not detect any representative of Pinguiophyceae, a class which include the LCD-producer *Phaeomonas parva* (Table S1). Reads associated to Pelagophyceae, and mostly (97%) affiliated to *Pelagomonas calceolata*, were recovered more frequently as they were present in 55 samples with an average abundance of 85 reads (2% of total reads) per sample and a maximum value of 935 reads (12% of total) in the DCM of Station 23 (Fig. 4B). Pelagophyceae reads were mostly detected in the DCM and were particularly abundant at the 3 westernmost stations investigated, where they comprised 8% of total reads (Fig. 4B).

Chrysophyceae and Dictyochophyceae were also detected in most samples (54 and 57 samples, respectively) and their reads were recovered more frequently at the surface and BWML of the westernmost part of the transect (Stations 20-23) and at the surface of Stations

393 3-4 (Fig. 4C and D). Their 18S rRNA gene reads reached abundances of up to 55 and 41 reads
394 (0.4 and 0.6% of total, respectively), for Chrysophyceae and Dictyochophyceae respectively,
395 in the BWML of Station 20 (Table S4). Raphidophyceae were present only in three samples
396 from Stations 11, 12, and 13 (Fig. 4F).
397
398

4. DISCUSSION

4.1. Comparison of diol distributions

In general, it is thought that 1,13- and 1,15-diols derive from a different source than 1,14-diols in the marine realm (Sinninghe Damsté et al., 2003; Rampen et al., 2007; Rampen et al., 2011). Indeed, linear regressions showed that the concentration of C_{30} 1,15-diol is significantly correlated with those of the C_{30} 1,13- and C_{32} 1,15-diols (Figs. 5A-B). We did not observe any significant correlation between the concentrations of the C_{28} 1,13- and the C_{30} 1,13- or C_{30} 1,15-diol (Figs. 5C-D), which might be due to the fact that C_{28} 1,13-diol was below detection limit in 19 out of 71 samples and its distribution could be compared to that of the widespread $C_{30\text{-}32}$ diols only for the remaining 52 samples. This low abundance of C_{28} 1,13-diol is consistent with the relatively high temperatures observed for the tropical Atlantic ocean (Fig. 1B), since the LCD core top calibration study has revealed that the fractional abundance of the C_{30} 1,15-diol is high and that of the C_{28} 1,13-diol is low when SST is relatively high (Rampen et al., 2012).

The concentration of the C_{28} 1,14-diol was not correlated with that of the C_{30} 1,14-diol (Fig. 5E), potentially suggesting a different origin for the C_{28} and the C_{30} 1,14-diols. However, the concentration of C_{30} 1,14-diol was significantly correlated with the C_{30} 1,15-diol (Fig. 5F). This is quite surprising as the 1,14-diols in seawater have been suggested to derive from *Proboscia* spp. (Sinninghe Damsté et al., 2003; Rampen et al., 2009), and to a lesser extent from *A. radians* (Rampen et al., 2011), whereas the 1,13- and 1,15-diols are thought to be associated with Eustigmatophyceae (Rampen et al., 2014 and references cited therein). Previous studies highlighted indeed good correlations in the fluxes of C_{28} and C_{30} 1,14-diols in the water column of the Arabian Sea (Rampen et al., 2007) and the northwestern Indian Ocean (Rampen et al., 2008). *Proboscia* spp. contain also unsaturated 1,14-diols which were not found here; specifically the warm water species *Proboscia indica* is dominated by $C_{28:1}$ and $C_{30:1}$ 1,14-diols (Rampen et al., 2007), suggesting that the 1,14-diols found here do not derive from *Proboscia* spp.. This is confirmed by the absence or very low proportions of 18S rRNA gene reads from the major producers of $C_{28\text{-}30}$ 1,14-diols, that are *Proboscia* spp. and *A. radians* (Table 1). This suggests different sources for the C_{28} and the C_{30} 1,14-diols. Since the C_{30} 1,15-diol accounted for >95% of the $C_{28\text{-}32}$ diols, it is possible that the C_{30} 1,14-diol was biosynthesised in low amounts, along with C_{30} 1,13-diol, by the producers of C_{30} 1,15-diol. This is supported by the fact that Eustigmatophyceae can contain small amounts of 1,14-diols along with large quantities of 1,15-diols (Rampen et al., 2014a); specifically the

C_{28} 1,14-diol, accounts for up to the 15% of the total LCDs in *Pseudostarastrum enorme*, and lower proportions (1-5%) of C_{30} 1,14-diols were previously found in *Vischeria punctata* and *Eustigmatos vischeri* (Rampen et al., 2014a).

It has been reported that the distributions of LCDs can be affected by riverine input, which is reflected by elevated amounts of the C_{32} 1,15-diol (>10%, de Bar et al., 2016; Lattaud et al., 2017b). However, the fractional abundance of the C_{32} 1,15-diol in the SPM is low (0 to 4%, data not shown), far lower than the values typically measured in river-influenced ecosystems such as the Iberian Atlantic Margin (de Bar et al., 2016), the Kara Sea (Lattaud et al., 2017b) or the Congo River plume (Versteegh et al., 2000). We did not detect other eustigmatophycean biomarkers such as C_{32} alkenols or C_{30-32} hydroxy fatty acids (Volkman et al., 1992; Gelin et al., 1997b), suggesting that riverine or marine Eustigmatophyceae were unlikely to source the C_{28-32} diols found here. The HCC cruise took place in a period of the year (August/September) when the water discharge from the Amazon River is typically low (Molleri et al., 2010), thus leading to low inputs of riverine organic matter into the sea. The distribution of LCDs in the sampled SPM is thus likely not impacted by terrestrial input of LCDs.

Beyond Heterokontophyta, LCDs may also be produced by lower (Speelman et al., 2009) and higher (Wen and Jetter, 2007; Racovita and Jetter, 2016) plants. However, only 4 reads from our dataset were associated with a plant species, i.e. *Panax ginseng* (Table S4), which is not known to contain LCDs. The near absence of 18S rRNA gene reads from higher plants confirms the low riverine input of organic matter in the SPM of the tropical North Atlantic waters analysed here.

We explored the variations in the concentrations of the LCDs with respect to environmental data. The C_{28} 1,13- and 1,14-diols, both occurring in low abundance, did not exhibit significant correlations with any of the environmental data measured here (Table 2). In contrast the concentrations of C_{30} 1,13-, 1,14- and 1,15-diols exhibited significant but weak positive correlations with temperature and dissolved silica and weak negative correlations with salinity and nitrite. The concentration of the C_{32} 1,15-diol revealed a correlation with the same environmental variables as the C_{30} diols except for dissolved silica and nitrite and exhibited a weak negative correlation with the concentration of nitrate. The correlations found here are likely simply due to different water masses: the mixed layer, where the highest proportions of LCDs were measured, exhibited indeed higher temperatures and lower salinities compared to the DCM. We repeated the analyses after excluding DCM samples and did not find strong positive or negative correlations between LCDs and environmental

variables (data not shown). Thus, there does not seem to be a major control of environmental conditions on the concentrations of LCDs.

4.2 Comparison with eukaryotic abundance and diversity

Although C_{28-32} diols are likely produced by phytoplankton, the variability in LCD abundance is not correlated with that of Chl-*a* concentration, or photosynthetic pico- and nanoeukaryote abundances (Table 2). This lack of correlation suggests that the LCD-producers accounted for only a small proportion of phytoplankton. The high proportion of Dinophyceae, Syndiniales, and Radiolaria revealed by our genetic libraries agree with previous studies on marine microbial communities based on 18S rRNA gene sequencing in different environments (Comeau et al., 2011; Christaki et al., 2014; de Vargas et al., 2015). However, these taxa do not necessarily dominate marine microbial communities and so our results are likely due to a relatively high number of rRNA gene copies per cell (Zhu et al., 2005). Larger-sized dinoflagellates such as *Prorocentrum minimum* and *Amphidinium carterae* can contain up to 1000 gene copies per cell compared to <10 of rRNA gene copies for smaller sized (<3 µm) species of Chlorophyta, Pelagophyceae, and Haptophyta (Zhu et al., 2005).

4.2.1 LCD-producers

Although the primers used in this study have a perfect match with the 18S rRNA gene sequences of most eukaryotes (including all the classes containing LCD-producers), and the rarefaction curves indicate that we sampled an appropriate (i.e. >90%) proportion of the eukaryotic community, we cannot fully exclude that some species remain undetected because of under sampling, or primer mismatches. Moreover, the large number (100-1000) of rRNA gene copies per cell present within dinoflagellates and Radiolaria might have somehow affected the detection of LCD producers. In particular *Nannochloropsis salina* has been shown to possess only 1-2 copies of 18S rRNA gene (Zhu et al., 2005), and similarly, the other marine *Nannochloropsis* species, which do not differ greatly in size from *N. salina* (Fawley and Fawley, 2007), are also likely to have a low number of 18S rRNA gene copies. Known species of LCD-producers were present in only 51% of our SPM samples as revealed by sequencing data (Table 1), whereas the major LCD, the C_{30} 1,15-diol, was present in all samples. This suggests that the LCDs found here were (1) either produced by other species which were not detected using the current methodology, or (2) that the LCD-producers were under sampled because of their low number of rRNA gene copies per cell, or (3) that the

DNA of the LCD-producers was no longer present in the SPM at the moment of sampling. Specifically, marine Eustigmatophyceae were represented by only two OTUs (denovo2075, *Nannochloropsis oculata*, and denovo229, uncultured Eustigmatophycea, Table S4) detected in only 8 samples, confirming the hypothesis of Volkman et al. (1992) and Rampen et al. (2012) that they are not the major producers of LCDs in the marine environment. Even if we expand our analyses of LCD-related species to a higher taxonomic level, we do not find large proportions of 18S rRNA reads (generally <0.9-% of total reads) except for the class Pelagophyceae, which accounts for up to 12-% of total reads (Fig. 2A-E). However, Pelagophyceae are unlikely to be the source of any of the LCDs found here because their vertical distribution (i.e. mostly detected in the DCM, Figs. 3, 4B) does not correspond well to that of LCDs, which were either more abundant in the upper layers (C_{30} 1,13-, 1,14-, and 1,15-diols and C_{32} 1,15-diol) or did not vary greatly with depth (C_{28} diols, Fig. 2). Chrysophyceae and Dictyochophyceae were instead more abundant in the upper layers (Fig. 4B-C) and although none of the three known LCD-producers from these classes produces the most abundant LCD detected in the SPM, i.e. C_{30} 1,15-diol (Table S1), other species within the Chrysophyceae and Dictyochophyceae may possibly be a source for the C_{30} diols.

The C_{28} diols exhibited higher concentrations at the BWML of Station 12 and at surface in Station 13 (Fig. 2A and B), and higher proportions of 18S rRNA gene reads were recovered from Pelagophyceae (2.4%), and Eustigmatophyceae (0.5%), at the surface of Stations 11-12 (Fig. 4D-F). The scattered occurrences of these groups and the mismatches in distributions when compared to the LCDs suggest that the LCDs in the tropical North Atlantic Ocean are unlikely to derive from Pelagophyceae, radial centric diatoms, Raphidophyceae, and/or Eustigmatophyceae.

Overall the abundance of known LCD producers is low and scattered and does not match the observed abundance patterns observed for the LCDs, suggesting that most of the LCDs measured here were not produced by any of these species.

4.2.2 Correlations between the abundance of OTUs and LCD concentration

Since LCDs have been shown to be present within two genetically distant eukaryotic supergroups, the Heterokontophyta and the Archaeplastida, the latter including plants as well as green and red algae, the genetic and enzymatic machinery required for the biosynthesis of LCDs might be present in other genera and classes, including uncultured species. We, therefore, also compared the concentration of LCDs with the composition of the entire eukaryotic microbial community, normalised with respect to the 18S rRNA gene abundance,

at both class and OTU levels to identify co-occurrence patterns. No significant correlation was found at class level (data not shown), whereas the correlations at the OTU level were weak ($r \leq 0.60$) but significant (p-value <0.01) for 27 OTUs affiliated to 11 different classes (Table 2). A reason behind the lack of correlation between taxonomic classes and LCDs can be that pooling OTUs at higher taxonomic levels likely leads to combining the LCD-producers with species which are unable to produce LCDs but that are falling in the same taxonomic level. The ability of microorganisms to biosynthesize LCDs can indeed vary, even between genetically related species; some genera include both LCD-producers and species which do not contain LCDs (Table S1).

The C_{30} 1,15-diol exhibited significant correlations (p <0.01) with 23 OTUs and overall, 27 OTUs were significantly correlated with C_{30} or to a lesser extent, C_{32} diols (Table 3). Of the 27 OTUs, 4 OTUs were affiliated to classes containing known LCD-producers (Chrysophyceae and Dictyochophyceae, Table 3). The abundance of the two chrysophycean OTUs (denovo465 and denovo1680, Table 3) exhibited significant correlations with the concentrations of both C_{30} 1,13- and 1,15-diols and accounted for 52% of the total reads from this class and the only known LCD-producer from this class (*Chrysosphaera parvula*) was found to contain C_{32} 1,15-diol (Rampen, unpublished results). The two OTUs affiliated to Dictyochophyceae (denovo873 and denovo958) and exhibiting positive correlation with $C_{30\text{-}32}$ diols, cluster within Pedinellales and Florenciellales families, respectively, and are thus closely related to two known LCD-producers, *Florenciella parvula* and *Apedinella radians*. However, *F. parvula* contains C_{24} 1,13-, C_{24} 1,14-, and C_{24} 1,15-diols (Rampen, unpublished results) and *A. radians* produces C_{28}, C_{30}, and C_{32} 1,14-diols (Rampen et al., 2011), whereas the two dictyochophycean OTUs denovo873 and denovo958 exhibited positive correlation with the C_{30} 1,15-diol (Table 3).

The correlation values found here are nearly all low ($r \approx 0.4\text{-}0.5$), raising the question of whether these relationships reflect the ability of these species to produce LCDs or whether they are simply driven by other environmental conditions leading to similar spatial distributions of OTUs and LCDs. Other OTUs showing significant correlations with C_{30} 1,15-diols are rare in the marine environment. For example, species falling in the Centroheliozoa (OTU denovo1066) are mostly known as freshwater predators (Slapeta et al., 2005). In seawater, they have only been sporadically detected in anoxic environments (Stock et al., 2009; Stoeck et al., 2009), suggesting that the centroheliozoan reads found here are unlikely to derive from active microorganisms. In contrast, the other OTUs include marine representatives commonly found in the photic zone of seawater and thus the reads found here

might derive from living organisms: Syndiniales are intracellular parasites of other marine protists, and the genetic clades found here (Group I Clade 4, Group II Clades 2, 7, 8, 17, and 23) are commonly detected in the upper 100 m of the water column (Guillou et al., 2008). Spirotrichea include several heterotrophic and mixotrophic marine planktonic ciliates (Agatha et al., 2004; Santoferrara et al., 2017), whereas *Phaeocystis* is a widespread primary producer. The OTUs of uncultured classes exhibiting significant positive correlations with LCDs (Prasino Clade IX and the HAP-3 clade) are also commonly observed in the photic zone (Shi et al., 2009; Egge et al., 2015; Lopes dos Santos et al., 2016). However, cultivated representatives would be required in order to confirm whether species within these clades are capable of LCD synthesis.

4.3 Can 18S rRNA gene-based community composition analysis be used to determine LCD biological sources?

The lack of correlations of C_{28} diols with any OTUs as well as the low degree of correlation between OTUs and $C_{30\text{-}32}$ diols and the trace abundance or near absence of known LCD producers suggest that the 18S rRNA genes from the microorganisms sourcing the LCDs were either absent, or present below detection level in the seawater sampled. The fact that we sampled >90% of the OTUs potentially present (Figure S1) and the use of universal eukaryotic primers suggests that LCD-producers have been unlikely to escape detection. However, the relatively low number of rRNA gene copies found for *N. oculata* (Zhu et al., 2005) and likewise also in other smaller-sized marine Eustigmatophyceae, suggest that LCD-producers might have been under sampled with respect to larger-sized species which can contain up to 1000 rRNA copies per cell (Zhu et al., 2005).

It should be considered that both the LCDs and DNA in the SPM might derive not only from active or senescent cells, but also from detritus (Not et al., 2009). In addition, LCDs can persist in seawater for likely much longer periods than the DNA of the related LCD-producers. Although the biological function of LCDs is unclear for most species, they have been shown to be the building blocks of cell wall polymers in Eustigmatophyceae, and likewise they might occur in other biopolymers of marine or terrestrial origin. In *Nannochloropsis* cell wall, LCDs and long chain alkenols are likely to be bound together through ester and ether bonds to form highly refractory polymers known as algaenans (Gelin et al., 1997a; Scholz et al., 2014). These biopolymers are thought to be quite persistent and accumulate in ancient sediments for millions of years (Tegelaar et al., 1989; Derenne and Largeau, 2001; de Leeuw et al., 2006). Indeed, LCDs are ubiquitous in recent surface

sediments (Rampen et al., 2012) and ancient sediments of up to 65 million years old (Yamamoto et al., 1996) showing their recalcitrant nature.

Recent laboratory experiments highlighted that LCDs from dead biomass of *Nannochloropsis oculata* can persist in seawater for longer than 250 days under both anoxic (Grossi et al., 2001) and oxic conditions (Reiche et al., in press). In contrast, much shorter turnover times (6 h to 2 months) are typically reported for extracellular DNA in the oxic water column (Nielsen et al., 2007). This suggests that the DNA from LCD-producers likely reflects the living eukaryotic community (recently) present when seawater was sampled, while the LCDs probably represent an accumulation that occurred over longer periods of time (weeks to months or even years).

Because of this large difference in turnover rates between LCDs and the DNA from the LCD-producers, 18S rRNA gene analysis of environmental samples may be unsuccessful for identifying LCD-producers. This is seemingly in contrast to a previous study that showed that the LCD concentration in the upper 25 m of the freshwater lake Challa (Tanzania) was related to the number of eustigmatophycean 18S rRNA gene copies (Villanueva et al., 2014). However, Villanueva et al. (2014) used Eustigmatophyceae-biased primers and since this was a lake system, Eustigmatophyceae are likely to be the major source of LCDs in freshwater ecosystems. Importantly, they found a mismatch for the uppermost part of the water column (0–5 m), where high LCD abundance (38–46 ng L^{-1}) coincided with little or no Eustigmatophyceae 18S rRNA gene copies. This pattern was explained by them to be caused by wind-driven and convective mixing of preserved LCDs, while phytoplankton adjusted its buoyancy at greater depth (Villanueva et al., 2014). The high salinity values (≥33 g kg^{-1}) detected in most surface samples, the low proportions of both C_{32} 1,15--diols (2.2% over the total LCDs) and 18S rRNA gene reads associated with plants (4 out of 238 564), as well as the low input of freshwater from the Amazon River to the stations analysed here during the sampling period (Molleri et al., 2010) suggest that the LCDs found here are unlikely to have a freshwater origin.

Laboratory experiments carried out under different conditions of temperature, light irradiance, salinity and nitrate concentrations revealed average cellular LCD content of about 23 fg $cell^{-1}$ (Balzano et al., 2017) for *Nannochloropsis oceanica*. The average LCD concentration in the SPM investigated was ca. 2.6 ng L^{-1}, which would correspond to ca. 1.1 · 10^6 pico/nano algal cells L^{-1}. We detected average phytoplankton abundances of 3.3 · 10^6 cell L^{-1} for picoeukaryotes and 3.6 · 10^4 cell L^{-1} for nanoeukaryotes. Although nanoplanktonic Eustigmatophyceae might produce larger amounts of LCDs than those measured in our

previous study (Balzano et al., 2017), because of their larger cell size, the nanoplankton abundances measured here are two orders of magnitude lower than the densities required to source the LCDs ($1.1 \cdot 10^6$ cell L^{-1}). Therefore, if the LCDs measured here were biosynthesised by intact microorganisms in the water column, nanoplankton alone would not be able to source all the LCDs measured, and therefore in addition at least one-third of the picophytoplankton should be able to produce LCDs, which is unrealistic. This supports the idea that most of the LCDs detected here are of fossil nature and not contained in living cells. The higher concentrations of LCDs found in the SPM from the mixed layer compared to the DCM suggest that LCDs were originally produced at a higher frequency in the mixed layer. Moreover, their possible fossil nature indicates that LCDs were likely to persist in the mixed layer for long periods, eventually associated with suspended particulate matter.

The combination of lipid and DNA analyses is often complicated by different turnover rates, especially for refractory compounds such as LCDs. Studies focused on more labile biomarker lipids such as fatty acids or intact polar lipids can be more successful, e.g. with short branched fatty acids (Balzano et al., 2011), cyanobacterial glycolipids (Bale et al., 2018), or archaeal phospholipids (Pitcher et al., 2011; Buckles et al., 2013). Therefore, care has to be taken in inferring sources of biomarker lipids by the quantitative comparison of DNA abundance with biomarker lipid concentrations. Analysis of intact polar lipids, rather than total lipids, might have facilitated the identification of diol producers.

5. Conclusions

The combination of lipid analyses and 18S rRNA gene amplicon sequencing revealed some weak correlations between the abundances of 27 OTUs and the concentration of C_{30} diols. Four of these OTUs are affiliated to classes that include few LCD-producing species (i.e. Chrysophyceae and Dictyochophyceae), whereas the remaining 23 OTUs belong to taxa in which the presence of LCDs has never been assessed. In both cases it remains unclear whether the correlation between these 27 OTUs and the C_{30} diols reflects novel LCD-producers or is driven by other environmental conditions.

The abundances of photosynthetic pico and nanoeukaryotes measured here suggest that these microbial populations are highly unlikely to source all the LCDs found. Some of the LCDs found here might be associated with suspended debris rather than intact cells, with the DNA from their producers being already degraded at the time of sampling. DNA degradation rates in the oxygenated water column are indeed faster than those of most lipids, including LCDs. The freshness of the organic matter and the turnover rates of both lipids and DNA in a given environment should thus be considered when identifying the biological sources of a specific class of lipids through DNA sequencing. In addition, the extraction methods applied in our study did not discriminate between free and bound lipids and we thus do not know if the compounds found here were originally present in seawater as free or ester-bound diols. Finally, the 18S rRNA gene amplicon sequencing can be suitable to track LCD sources (1) for simple ecosystems or laboratory/*in situ* mesocosms with high proportions of fresh organic matter and (2) for low oxygen/anoxic environments where extracellular DNA can persist for longer periods.

Acknowledgments

–We thank the captains and the crew of the R/V Pelagia for their support during the cruise. We thank Denise Dorhout for sample collection, Sharyn Ossebaar for nutrient analysis, H. Witte, E. Panoto, and S. Vreugdenhil for support in molecular biology, H. Malschaert for the bioinformatics and M. Besseling for helpful discussions. John Volkman and two anonymous referees provided helpful comments on an earlier version of this paper. This research was funded by the European Research Council (ERC) under the European Union's Seventh Framework Program (FP7/2007-2013) ERC grant agreement [339206]. S.S. and J.S.S.D. receive financial support from the Netherlands Earth System Science Centre (NESSC) though a gravitation grant (NWO 024.002.001) from the Dutch Ministry for Education, Culture and Science.

References

Agatha, S., Struder-Kypke, M. C., and Beran, A.: Morphologic and genetic variability in the marine planktonic ciliate *Laboea strobila* Lohmann, 1908 (Ciliophora, Oligotrichia), with notes on its ontogenesis, J. Eukaryot. Microbiol., 51, 267-281, https://doi.org/10.1111/j.1550-7408.2004.tb00567.x, 2004.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J.: Basic local alignment search tool, J. Mol. Bio., 215, 403-410, http://dx.doi.org/10.1016/S0022-2836(05)80360-2, 1990.

Andersen, R. A., Brett, R. W., Potter, D., and Sexton, J. P.: Phylogeny of the Eustigmatophyceae based upon 18S rDNA, with emphasis on *Nannochloropsis*, Protist, 149, 61-74, https://doi.org/10.1016/S1434-4610(98)70010-0, 1998.

Bale, N. J., Villareal, T. A., Hopmans, E. C., Brussaard, C. P. D., Besseling, M., Dorhout, D., Sinninghe Damsté, J. S., and Schouten, S.: C5 glycolipids of heterocystous cyanobacteria track symbiont abundance in the diatom *Hemiaulus hauckii* across the tropical north Atlantic, Biogeosciences, 15, 1229–1241, https://doi.org/10.5194/bg-2017-300, 2018.

Balzano, S., Pancost, R. D., Lloyd, J. R., and Statham, P. J.: Changes in fatty acid composition in degrading algal aggregates, Mar. Chem., 124, 2-13, https://doi.org/10.1016/j.marchem.2010.11.001, 2011.

Balzano, S., Abs, E., and Leterme, S. C.: Protist diversity along a salinity gradient in a coastal lagoon, Aquat. Microb. Ecol., 74, 263-277, https://doi.org/10.3354/ame01740, 2015.

Balzano, S., Villanueva, L., de Bar, M., Sinninghe Damsté, J. S., and Schouten, S.: Impact of culturing conditions on the abundance and composition of long chain alkyl diols in species of the genus *Nannochloropsis*, Org. Geochem., 108, 9-17, https://doi.org/10.1016/j.orggeochem.2017.02.006, 2017.

Benjamini, Y., and Hochberg, Y.: Controlling the false discovery rate. A pratical and powerful approach to multiple testing., J Roy. Stat. Soc. B Met., 57, 289-300, 1995.

Bergesch, M., Odebrecht, C., and Moestrup, O.: Nanoflagellates from coastal waters of southern Brazil (32 degrees S), Bot. Mar., 51, 35-50, https://doi.org/10.1515/bot.2008.003, 2008.

Buckles, L. K., Villanueva, L., Weijers, J. W. H., Verschuren, D., and Sinninghe Damste, J. S.: Linking isoprenoidal GDGT membrane lipid distributions with gene abundances of ammonia-oxidizing Thaumarchaeota and uncultured crenarchaeotal groups in the water column of a tropical lake (Lake Challa, East Africa), Environ. Microbiol., 15, 2445-2462, https://doi.org/10.1111/1462-2920.12118, 2013.

Buschhaus, C., Peng, C., and Jetter, R.: Very-long-chain 1,2-and 1,3-bifunctional compounds from the cuticular wax of Cosmos bipinnatus petals, Phytochemistry, 91, 249-256, https://doi.org/10.1016/j.phytochem.2012.07.018, 2013.

731 Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N.,
732 Pena, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E.,
733 Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J.
734 R., Tumbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R.:
735 QIIME allows analysis of high-throughput community sequencing data, Nat. Methods, 7, 335-
736 336, https://doi.org/10.1038/nmeth.f.303, 2010.

737 Christaki, U., Kormas, K. A., Genitsaris, S., Georges, C., Sime-Ngando, T., Viscogliosi, E., and Monchy,
738 S.: Winter-summer succession of unicellular eukaryotes in a meso-eutrophic coastal system,
739 Microb. Ecol., 67, 13-23, https://doi.org/10.1007/s00248-013-0290-4, 2014.

740 Comeau, A. M., Li, W. K. W., Tremblay, J.-E., Carmack, E. C., and Lovejoy, C.: Arctic Ocean microbial
741 community structure before and after the 2007 record sea ice minimum, Plos One, 6,
742 10.1371/journal.pone.0027492, https://doi.org/10.1371/journal.pone.0027492, 2011.

743 de Bar, M. W., Dorhout, D. J. C., Hopmans, E. C., Rampen, S. W., Sinninghe Damsté, J. S., and
744 Schouten, S.: Constraints on the application of long chain diol proxies in the Iberian Atlantic
745 margin, Org. Geochem., 101, 184-195, https://doi.org/10.1016/j.orggeochem.2016.09.005,
746 2016.

747 de Leeuw, J. W., Irene, W., Rijpstra, C., and Schenck, P. A.: The occurrence and identification of C_{30},
748 C_{31} and C_{32} alkan-1, 15-diols and alkan-15-one-1-ols in Unit I and Unit II Black Sea sediments,
749 Geochim. Cosmochim. Ac., 45, 2281-2285, http://dx.doi.org/10.1016/0016-7037(81)90077-6,
750 1981.

751 de Leeuw, J. W., Versteegh, G. J. M., and van Bergen, P. F.: Biomacromolecules of algae and plants
752 and their fossil analogues, Plant Ecol., 182, 209-233, https://doi.org/10.1007/s11258-005-
753 9027-x, 2006.

754 de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N.,
755 Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron,
756 S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horák, A., Jaillon, O., Lima-Mendez, G.,
757 Lukeš, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A.,
758 Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Acinas, S. G., Bork, P., Bowler, C.,
759 Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J.,
760 Sieracki, M. E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., and
761 Karsenti, E.: Eukaryotic plankton diversity in the sunlit ocean, Science, 348,
762 https://doi.org/10.1126/science.1261605, 10.1126/science.1261605, 2015.

763 Derenne, S., and Largeau, C.: A review of some important families of refractory macromolecules:
764 Composition, origin, and fate in soils and sediments, Soil Sci., 166, 833-847,
765 https://doi.org/10.1097/00010694-200111000-00008, 2001.

766 Dixon, P.: VEGAN, a package of R functions for community ecology, J. Veg. Sci., 14, 927-930,
767 https://doi.org/10.1111/j.1654-1103.2003.tb02228.x, 2003.

768 Edgar, R. C.: Search and clustering orders of magnitude faster than BLAST, Bioinformatics, 26, 2460-
769 2461, https://doi.org/10.1093/bioinformatics/btq461, 2010.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R.: UCHIME improves sensitivity and speed of chimera detection, Bioinformatics, 27, 2194-2200, https://doi.org/10.1093/bioinformatics/btr381, 2011.

Egge, E. S., Johannessen, T. V., Andersen, T., Eikrem, W., Bittner, L., Larsen, A., Sandaa, R.-A., and Edvardsen, B.: Seasonal diversity and dynamics of haptophytes in the Skagerrak, Norway, explored by high-throughput sequencing, Mol. Ecol., 24, 3026-3042, https://doi.org/10.1111/mec.13160, 2015.

Fawley, K. P., and Fawley, M. W.: Observations on the diversity and ecology of freshwater *Nannochloropsis* (Eustigmatophyceae), with descriptions of new taxa, Protist, 158, 325-336, https://doi.org/10.1016/j.protis.2007.03.003, 2007.

Gelin, F., Boogers, I., Noordeloos, A. A. M., Sinninghe Damsté, J. S., Riegman, R., and De Leeuw, J. W.: Resistant biomacromolecules in marine microalgae of the classes Eustigmatophyceae and Chlorophyceae: geochemical implications, Org. Geochem., 26, 659-675, https://doi.org/10.1016/s0146-6380(97)00035-1, 1997a.

Gelin, F., Volkman, J. K., deLeeuw, J. W., and Sinninghe Damsté, J. S.: Mid-chain hydroxy long-chain fatty acids in microalgae from the genus *Nannochloropsis*, Phytochemistry, 45, 641-646, https://doi.org/10.1016/s0031-9422(97)00068-x, 1997b.

Grossi, V., Blokker, P., and Sinninghe Damste, J. S.: Anaerobic biodegradation of lipids of the marine microalga *Nannochloropsis salina*, Org. Geochem., 32, 795-808, https://doi.org/10.1016/s0146-6380(01)00040-7, 2001.

Guillou, L., Viprey, M., Chambouvet, A., Welsh, R. M., Kirkham, A. R., Massana, R., Scanlan, D. J., and Worden, A. Z.: Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata), Environ. Microbiol., 10, 3349-3365, https://doi.org/10.1111/j.1462-2920.2008.01731.x, 2008.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., Mahé, F., Massana, R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A.-L., Siano, R., Stoeck, T., Vaulot, D., Zimmermann, P., and Christen, R.: The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy, Nucleic Acid Res., 41, 597-604, https://doi.org/10.1093/nar/gks1160, 2013.

Jetter, R., and Riederer, M.: Long-chain alkanediols, ketoaldehydes, ketoalcohols and ketoalkyl esters in the cuticular waxes of Osmunda regalis fronds, Phytochemistry, 52, 907-915, https://doi.org/10.1016/s0031-9422(99)00309-x, 1999.

Jiang, S. C., Oleary, T., Volkman, J. K., Zhang, H. Z., Jia, R. F., Yu, S. H., Wang, Y., Luan, Z. F., Sun, Z. Q., and Jiang, R. H.: Origins and simulated thermal alteration of sterols and keto-alcohols in deep-sea marine sediment of the Okinawa Trough., Org. Geochem., 21, 415-422, https://doi.org/10.1016/0146-6380(94)90203-8, 1994.

Koning, E., van Iperen, J. M., van Raaphorst, W., Helder, W., Brummer, G. J. A., and van Weering, T. C. E.: Selective preservation of upwelling-indicating diatoms in sediments off Somalia, NW Indian Ocean, Deep-Sea Research Part I-Oceanographic Research Papers, 48, 2473-2495, https://doi.org/10.1016/s0967-0637(01)00019-x, 2001.

Lassiter, A. M., Wilkerson, F. P., Dugdale, R. C., and Hogue, V. E.: Phytoplankton assemblages in the CoOP-WEST coastal upwelling area, Deep-Sea Research Part II-Topical Studies in Oceanography, 53, 3063-3077, https://doi.org/10.1016/j.dsr2.2006.07.013, 2006.

Lattaud, J., Dorhout, D., Schulz, H., Castañeda, I. S., Schefuß, E., Sinninghe Damsté, J. S., and Schouten, S.: The C_{32} alkane-1,15-diol as a proxy of late Quaternary riverine input in coastal margins, Clim. Past, 13, 1049-1061, https://doi.org/10.5194/cp-13-1049-2017, 2017a.

Lattaud, J., Kim, J. H., De Jonge, C., Zell, C., Sinninghe Damste, J. S., and Schouten, S.: The C_{32} alkane-1,15-diol as a tracer for riverine input in coastal seas, Geochim. Cosmochim. Ac., 202, 146-158, https://doi.org/10.1016/j.gca.2016.12.030, 2017b.

Legendre, P., and Gallagher, E. D.: Ecologically meaningful transformations for ordination of species data, Oecologia, 129, 271-280, https://doi.org/10.1007/s004420100716, 2001.

Logares, R., Audic, S., Santini, S., Pernice, M. C., de Vargas, C., and Massana, R.: Diversity patterns and activity of uncultured marine heterotrophic flagellates unveiled with pyrosequencing, ISME J, 6, 1823-1833, https://doi.org/10.1038/ismej.2012.36, 2012.

Lopes dos Santos, A., Gourvil, P., Tragin, M., Noel, M.-H., Decelle, J., Romac, S., and Vaulot, D.: Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae in oceanic waters, ISME J, 11, 512-528, https://doi.org/10.1038/ismej.2016.120, 2016.

Mao, S. Y., Zhu, X. W., Wu, N. Y., Jia, G. D., Sun, Y. G., Guan, H. X., and Wu, D. D.: Alcohol compounds in Azolla imbricata and potential source implication for marine sediments, Sci. China Earth Sci., 60, 348-359, https://doi.org/10.1007/s11430-016-5177-6, 2017.

Marie, D., Simon, N., and Vaulot, D.: Phytoplankton cell counting by flow cytometry., in: Algal Culturing Techniques, edited by: RA, A., Elsevier Academic Press: Burlington, MA, 253-268, 2005.

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., Chambouvet, A., Christen, R., Claverie, J. M., Decelle, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Forn, I., Forster, D., Guillou, L., Jaillon, O., Kooistra, W., Logares, R., Mahe, F., Not, F., Ogata, H., Pawlowski, J., Pernice, M. C., Probert, I., Romac, S., Richards, T., Santini, S., Shalchian-Tabrizi, K., Siano, R., Simon, N., Stoeck, T., Vaulot, D., Zingone, A., and de Vargas, C.: Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing, Environ. Microbiol., 17, 4035-4049, https://doi.org/10.1111/1462-2920.12955, 2015.

Moita, M. T., Oliveira, P. B., Mendes, J. C., and Palma, A. S.: Distribution of chlorophyll a and Gymnodinium catenatum associated with coastal upwelling plumes off central Portugal, Acta Oecol., 24, S125-S132, https://doi.org/10.1016/s1146-609x(03)00011-0, 2003.

Molleri, G. S. F., Novo, E., and Kampel, M.: Space-time variability of the Amazon River plume based on satellite ocean color, Cont. Shelf Res., 30, 342-352, https://doi.org/10.1016/j.csr.2009.11.015, 2010.

Nielsen, K. M., Johnsen, P. J., Bensasson, D., and Daffonchio, D.: Release and persistence of extracellular DNA in the environment, Environmental Biosafety Research, 6, 37-53, https://doi.org/10.1051/ebr:2007031, 2007.

Not, F., del Campo, J., Balague, V., de Vargas, C., and Massana, R.: New Insights into the Diversity of Marine Picoeukaryotes, Plos One, 4, 10.1371/journal.pone.0007143, https://doi.org/10.1371/journal.pone.0007143, 2009.

Pitcher, A., Villanueva, L., Hopmans, E. C., Schouten, S., Reichart, G. J., and Sinninghe Damste, J. S.: Niche segregation of ammonia-oxidizing archaea and anammox bacteria in the Arabian Sea oxygen minimum zone, ISME J, 5, 1896-1904, https://doi.org/10.1038/ismej.2011.60, 2011.

Plancq, J., Mattioli, E., Pittet, B., Simon, L., and Grossi, V.: Productivity and sea-surface temperature changes recorded during the late Eocene-early Oligocene at DSDP Site 511 (South Atlantic), Palaeogeogr. Palaeocl., 407, 34-44, https://doi.org/10.1016/j.palaeo.2014.04.016, 2014.

Racovita, R. C., and Jetter, R.: Identification of In-Chain-Functionalized Compounds and Methyl-Branched Alkanes in Cuticular Waxes of Triticum aestivum cv. Bethlehem, Plos One, 11, 25, https://doi.org/10.1371/journal.pone.0165827, 2016.

Rampen, S. W., Schouten, S., Wakeham, S. G., and Sinninghe Damste, J. S.: Seasonal and spatial variation in the sources and fluxes of long chain diols and mid-chain hydroxy methyl alkanoates in the Arabian Sea, Org. Geochem., 38, 165-179, https://doi.org/10.1016/j.orggepchem.2006.10.008, 2007.

Rampen, S. W., Schouten, S., Koning, E., Brummer, G.-J. A., and Sinninghe Damsté, J. S.: A 90 kyr upwelling record from the northwestern Indian Ocean using a novel long-chain diol index, Earth Planet. Sc. Lett., 276, 207-213, https://doi.org/10.1016/j.epsl.2008.09.022, 2008.

Rampen, S. W., Schouten, S., Schefuss, E., and Sinninghe Damsté, J. S.: Impact of temperature on long chain diol and mid-chain hydroxy methyl alkanoate composition in *Proboscia diatoms*: results from culture and field studies, Org. Geochem., 40, 1124-1131, https://doi.org/10.1016/j.orggeochem.2009.08.005, 2009.

Rampen, S. W., Schouten, S., and Sinninghe Damsté, J. S.: Occurrence of long chain 1,14-diols in *Apedinella radians*, Org. Geochem., 42, 572-574, https://doi.org/10.1016/j.orggeochem.2011.03.009, 2011.

Rampen, S. W., Willmott, V., Kim, J.-H., Uliana, E., Mollenhauer, G., Schefuss, E., Sinninghe Damsté, J. S., and Schouten, S.: Long chain 1,13-and 1,15-diols as a potential proxy for palaeotemperature reconstruction, Geochim. Cosmochim. Ac., 84, 204-216, https://doi.org/10.1016/j.gca.2012.01.024, 2012.

881 Rampen, S. W., Datema, M., Rodrigo-Gamiz, M., Schouten, S., Reichart, G.-J., and Sinninghe Damsté,
882 J. S.: Sources and proxy potential of long chain alkyl diols in lacustrine environments,
883 Geochim. Cosmochim. Ac., 144, 59-71, https://doi.org/10.1016/j.gca.2014.08.033, 2014a.

884 Rampen, S. W., Willmott, V., Kim, J.-H., Rodrigo-Gamiz, M., Uliana, E., Mollenhauer, G., Schefuss, E.,
885 Sinninghe Damsté, J. S., and Schouten, S.: Evaluation of long chain 1,14-alkyl diols in marine
886 sediments as indicators for upwelling and temperature, Org. Geochem., 76, 39-47,
887 https://doi.org/10.1016/j.orggeochem.2014.07.012, 2014b.

888 Reiche, S., Rampen, S., Dorhout, D., Sinninghe Damste, J. S., and Schouten, S.: The impact of oxygen
889 exposure on long-chain alkyl diols and the long chain diol index (LDI) – a long-term incubation
890 study, Biogeosciences, https://doi.org/10.1016/j.orggeochem.2018.08.003, in press.

891 Rodrigo-Gámiz, M., Rampen, S. W., de Haas, H., Baas, M., Schouten, S., and Sinninghe Damsté, J. S.:
892 Constraints on the applicability of the organic temperature proxies UK037, TEX86 and LDI in
893 the subpolar region around Iceland, Biogeosciences, 12, 6573-6590,
894 https://doi.org/10.5194/bg-12-6573-2015, 2015.

895 Santoferrara, L. F., Alder, V. V., and McManus, G. B.: Phylogeny, classification and diversity of
896 Choreotrichia and Oligotrichia (Ciliophora, Spirotrichea), Mol. Phylogenet. Evol., 112, 12-22,
897 https://doi.org/10.1016/j.ympev.2017.03.010, 2017.

898 Schlitzer, R.: Interactive analysis and visualization of geoscience data with Ocean Data View,
899 Computers & Geosciences, 28, 1211-1218, https://doi.org/10.1016/s0098-3004(02)00040-7,
900 2002.

901 Scholz, M. J., Weiss, T. L., Jinkerson, R. E., Jing, J., Roth, R., Goodenough, U., Posewitz, M. C., and
902 Gerken, H. G.: Ultrastructure and composition of the *Nannochloropsis gaditana* cell wall,
903 Eukaryot. Cell, 13, 1450-1464, https://doi.org/10.1128/ec.00183-14, 2014.

904 Seoane, S., Laza, A., Urrutxurtu, M., and Orive, E.: Phytoplankton assemblages and their dominant
905 pigments in the Nervion River estuary, Hydrobiologia, 549, 1-13,
906 https://doi.org/10.1007/s10750-005-1736-6, 2005.

907 Shi, X. L., Marie, D., Jardillier, L., Scanlan, D. J., and Vaulot, D.: Groups without Cultured
908 Representatives Dominate Eukaryotic Picophytoplankton in the Oligotrophic South East
909 Pacific Ocean, Plos One, 4, https://doi.org/10.1371/journal.pone.0007657, 2009.

910 Sinninghe Damsté, J. S., Rampen, S., Irene, W., Rupstra, C., Abbas, B., Muyzer, G., and Schouten, S.: A
911 diatomaceous origin for long-chain diols and mid-chain hydroxy methyl alkanoates widely
912 occurring in Quaternary marine sediments: Indicators for high-nutrient conditions, Geochim.
913 Cosmochim. Ac., 67, 1339-1348, https://doi.org/10.1016/s0016-7037(02)01225-5, 2003.

914 Slapeta, J., Moreira, D., and Lopez-Garcia, P.: The extent of protist diversity: insights from molecular
915 ecology of freshwater eukaryotes, P. Roy. Soc. B-Biol. Sci., 272, 2073-2081,
916 https://doi.org/10.1098/rspb.2005.3195, 2005.

Speelman, E. N., Reichart, G. J., de Leeuw, J. W., Rijpstra, W. I. C., and Sinninghe Damste, J. S.: Biomarker lipids of the freshwater fern Azolla and its fossil counterpart from the Eocene Arctic Ocean, Org. Geochem., 40, 628-637, https://doi.org/10.1016/j.orggeochem.2009.02.001, 2009.

Stock, A., Juergens, K., Bunge, J., and Stoeck, T.: Protistan diversity in suboxic and anoxic waters of the Gotland Deep (Baltic Sea) as revealed by 18S rRNA clone libraries, Aquat. Microb. Ecol., 55, 267-284, https://doi.org/10.3354/ame01301, 2009.

Stoeck, T., Behnke, A., Christen, R., Amaral-Zettler, L., Rodriguez-Mora, M. J., Chistoserdov, A., Orsi, W., and Edgcomb, V. P.: Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities, BMC Biol., 7, https://doi.org/10.1186/1741-7007-7-72, 2009.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D. M., Breiner, H. W., and Richards, T. A.: Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water, Mol. Ecol., 19, 21-31, https://doi.org/10.1111/j.1365-294X.2009.04480.x, 2010.

Tegelaar, E. W., Deleeuw, J. W., Derenne, S., and Largeau, C.: A reappraisal of kerogen formation, Geochim. Cosmochim. Ac., 53, 3103-3106, https://doi.org/10.1016/0016-7037(89)90191-9, 1989.

Versteegh, G. J. M., Bosch, H. J., and De Leeuw, J. W.: Potential palaeoenvironmental information of C-24 to C-36 mid-chain diols, keto-ols and mid-chain hydroxy fatty acids; a critical review, Org. Geochem., 27, 1-13, https://doi.org/10.1016/s0146-6380(97)00063-6, 1997.

Versteegh, G. J. M., Jansen, J. H. F., De Leeuw, J. W., and Schneider, R. R.: Mid-chain diols and keto-ols in SE Atlantic sediments: A new tool for tracing past sea surface water masses?, Geochim. Cosmochim. Ac., 64, 1879-1892, https://doi.org/10.1016/s0016-7037(99)00398-1, 2000.

Villanueva, L., Besseling, M., Rodrigo-Gamiz, M., Rampen, S. W., Verschuren, D., and Sinninghe Damsté, J. S.: Potential biological sources of long chain alkyl diols in a lacustrine system, Org. Geochem., 68, 27-30, https://doi.org/10.1016/j.orggeochem.2014.01.001, 2014.

Volkman, J. K., Barrett, S. M., Dunstan, G. A., and Jeffrey, S. W.: C-30-C-32 Alkyl diols and unsaturated alcohols in microalgae of the class Eustigmatophyceae., Org. Geochem., 18, 131-138, https://doi.org/10.1016/0146-6380(92)90150-v, 1992.

Wen, M., and Jetter, R.: Very-long-chain hydroxyaldehydes from the cuticular wax of Taxus baccata needles, Phytochemistry, 68, 2563-2569, https://doi.org/10.1016/j.phytochem.2007.05.029, 2007.

Yamamoto, M., Ficken, K., Baas, M., JH, B., and JW, d. L.: Molecular paleontology of the earliest danian at Geulhemmerberg (the Netherlands), Geol. Mijnbouw, 75, 255-267, 1996.

952 Zhu, F., Massana, R., Not, F., Marie, D., and Vaulot, D.: Mapping of picoeucaryotes in marine
953 ecosystems with quantitative PCR of the 18S rRNA gene, FEMS Microbiol. Ecol., 52, 79-92,
954 https://doi.org/10.1016/j.femsec.2004.10.006, 2005.

955

956

Table 1. Distribution of the18S rRNA gene reads associated with known LCD-producers

Taxon	Florenciellales	*Heterosigma*	Eustigmatophyceae	*Proboscia*	Total
No of samples[a]	28	2	8	2	35
Surface	12	1	2	0	12
BWML[b]	11	0	2	2	13
DCM[c]	5	1	4	0	10
No of reads[d]	99	10	45	3	157
% total	0.04	0.004	0.02	0.001	0.06
Surface	48	4	25	0	77
BWML	41	0	9	3	53
DCM	10	6	11	0	27

[a] Number of samples where 18S rRNA gene reads from C_{28-32} diol-producers were found Overall 68 samples were screened for the presence of 18S rRNA genes affiliated to LCD-producers.

[b] Bottom wind mixed layer

[c] Deep chlorophyll maximum

[d] Number or proportion of 18S rRNA gene reads associated with C_{28-32} diol-producers.

Table 2. Spearman rank correlation coefficients between LCD and environmental variables[a].

	C_{28} 1,13	C_{28} 1,14	C_{30} 1,13	C_{30} 1,14	C_{30} 1,15	C_{32} 1,15
Organic carbon	0.3	0.2	0.2	0.3	0.3	0.3
Salinity	-0.2	0.0	**-0.5**	**-0.7**	**-0.6**	**-0.6**
Temperature	0.2	-0.1	**0.5**	**0.5**	**0.5**	**0.5**
Phosphate	0.0	0.2	-0.3	-0.2	-0.3	-0.2
Ammonium	0.0	0.1	-0.3	-0.4	-0.4	-0.2
Nitrite	-0.2	0.0	**-0.6**	**-0.5**	**-0.6**	-0.4
Nitrate	0.0	0.2	-0.4	-0.3	-0.3	**-0.5**
Silica	0.1	0.0	0.4	**0.5**	**0.5**	0.4
Chl-*a*	-0.1	0.0	-0.2	-0.2	-0.3	-0.1
Picoeukaryotes	-0.1	-0.1	-0.4	-0.3	-0.4	-0.2
Nanoeukaryotes	0.0	-0.1	0.1	0.2	0.2	0.2

[a] Significant (p-value < 0.01) correlation values are in bold.

986 **Table 3. Correlation coefficient r** for the Operational Taxonomic Units (OTUs), representing 95 % of sequence identity, whose abundance was
987 correlated[a] with the concentration of LCDs in SPM samples obtained in the HCC cruise.
988
989

OTU ID[b]	**Taxon**	**Class**	**C_{30} 1,13**	**C_{30} 1,14**	**C_{30} 1,15**	**C_{32} 1,15**	**Total 1,13**	**Total 1,14**	**Total 1,15**
denovo2033	Choreotrichia	Spirotrichea							0.40
denovo2137	*Climacocylis scalaria*		0.45		0.49		0.45		0.49
denovo940	*Laboea strobila*		0.53	0.46	0.60		0.56	0.45	0.59
denovo685	Oligotrichia				0.41				0.40
denovo1804	*Pseudotontonia*		0.56	0.47	0.56	0.47	0.53	0.41	0.57
denovo492	*Blastodinium spinulosum*	Dinophyceae	0.43	0.44	0.46				0.45
denovo720	*Ceratocorys horrida*				0.46		0.44		0.45
denovo1682	*Neoceratium fusus*							0.47	
denovo526	*Protodinium simplex*			0.43	0.44			0.48	0.43
denovo267	*Pyrophacus_steinii*				0.43				0.42
denovo732	Dino Group I Clade 4	Syndiniales			0.40		0.46		0.41
denovo555	Dino Group II Clade 2				0.49		0.41		0.48
denovo1077	Dino Group II Clade 7				0.44		0.42		0.43
denovo1834	Dino Group II Clade 8				0.44		0.45		0.45
denovo1145	Dino Group II Clade 17		0.50		0.49		0.53	0.42	0.48
denovo2080	Dino Group II Clade 23				0.40		0.43		0.40
denovo725	Prasino Clade 9B	Prasino Clade IX			0.42		0.41		0.41
denovo1066	Pterocystida	Centroheliozoa			0.46				0.46
denovo400	HAP3	Haptophyta		0.47	0.49			0.47	0.48
denovo2132	*Phaeocystis*						0.46		
denovo972	*Haptolina*						0.44		
denovo465	**Chrysophyceae Clade G**	Chrysophyceae	0.44		0.43		0.45		0.42
denovo1680	**Chrysophyceae Clade H**		0.44		0.42		0.48		0.42
denovo1988	Raphid pennate	diatoms			0.41				0.41
denovo873	**Pedinellales**	Dictyochophyceae	0.56	0.45	0.55	0.52	0.55		0.56

denovo958	**Florenciellales**	Dictyochophyceae			0.43	0.45		0.44
denovo2433	Unidentified picozoan	Picozoa	0.49	0.50	0.55	0.47	0.46	0.55

[a] Only significant (p-value < 0.01 after FDR correction) correlations are shown.

[b] OTUs closely related to known LCD-producers are in bold.

997 **Figure Legend**

998

999 **Figure 1.** HCC cruise track in the western tropical North Atlantic Ocean, physical
1000 seawater properties, and biological parameters. (**A**) Map of the sampling stations. Spatial
1001 distribution of (**B**) temperature, (**C**) salinity, the concentration of (**D**) Chl-*a*, (**E**) organic
1002 carbon concentrations, and the abundance of photosynthetic (**F**) picoeukaryotes and (**G**)
1003 nanoeukaryotes. Temperature, salinity as well as the concentrations of Chl-*a* and organic
1004 carbon have been also published by Bale et al. (2018). Data were plotted using ODV software
1005 using kriging for interpolation between datapoints (Schlitzer, 2002). Dots represent the depth
1006 at which SPM was collected.

1007

1008 **Figure 2.** Spatial distribution of the concentration of LCDs: (**A**) C_{28} 1,13-diol, (**B**)
1009 C_{28} 1,14-diol, (**C**) C_{30} 1,13-diol, (**D**) C_{30} 1,14-diol, (**E**) C_{30} 1,15-diol, and (**F**) C_{32} 1,15-diol.
1010 Data were plotted using ODV software using kriging for interpolation between datapoints
1011 (Schlitzer, 2002).

1012

1013 **Figure 3.** Average fractional abundance of the reads obtained by 18S rRNA gene
1014 sequencing of SPM from the western tropical Atlantic Ocean over the various classes of
1015 eukaryotes. The V4 fragment of the 18S rRNA gene was sequenced using universal
1016 eukaryotic primers. Samples were pooled according to depth and the average contribution
1017 from each group at the different depth is shown. Error bars represent the standard deviation in
1018 the data from the various stations.

1019

1020 **Figure 4.** Spatial distribution of the 18S rRNA gene fragments related to taxa containing
1021 LCD-producers at different stations and depth. (**A**) Pelagophyceae, (**B**) Chrysophyceae, (**C**)
1022 Dictyochophyceae, (**D**) radial centric diatoms, (**E**) Eustigmatophyceae, and (**F**)
1023 Raphidophyceae. Data were plotted using ODV software using kriging for interpolation
1024 between datapoints (Schlitzer, 2002).

1025

Figure 5. Scatter plots of the concentrations of the different LCDs in the western tropical Atlantic Ocean. (**A**) C_{30} 1,13-diol vs C_{30} 1,15-diol, (**B**) C_{32} 1,15-diol vs C_{30} 1,15-diol, (**C**) C_{28} 1,13-diol vs C_{30} 1,13-diol, (**D**) C_{28} 1,13-diol vs C_{30} 1,15-diol, (**E**) C_{30} 1,14-diol vs C_{28} 1,14-diol, (**F**) C_{30} 1,14-diol vs C_{30} 1,15-diol.