Journal of
Plankton Research

plankt.oxfordjournals.org

Featured Article

J. Plankton Res. (2016) 38(5): 1135–1150. First published online June 29, 2016 doi:10.1093/plankt/fbw043

Phytoplankton assemblage during the North Atlantic spring bloom assessed from functional gene analysis

BESS B. WARD* AND NICOLAS VAN OOSTENDE
DEPARTMENT OF GEOSCIENCES, GUYOT HALL, PRINCETON UNIVERSITY, PRINCETON, NJ 08544, USA

*CORRESPONDING AUTHOR: bbw@princeton.edu

Received October 23, 2015; accepted May 25, 2016

Corresponding editor: Pia Moisander

The spring bloom in the North Atlantic develops over a few weeks in response to the physical stabilization of the nutrient-replete water column and is one of the biggest biological signals on earth. The composition of the phytoplankton assemblage during the spring bloom of 2008 was evaluated, using a microarray, on the basis of functional genes that encode key enzymes in nitrogen and carbon assimilation in eukaryotic and prokaryotic phytoplankton. The phytoarray is described, and its usefulness and limitations are demonstrated in this application to analysis of a spring bloom event. Oligonucleotide archetype probes representing ribulose bisphosphate carboxylase (RuBisCO), nitrate reductase and nitrate transporter genes from major phytoplankton classes detected a diverse assemblage. For RuBisCO, the archetypes with strongest signals represented known phytoplankton groups, but for the nitrate-related genes, the major signals were not closely related to any known phytoplankton sequences. Most of the assemblage's components exhibited consistent temporal/spatial patterns. Yet, the strongest archetype signals often showed quite different patterns, indicating different ecological responses by the main players. The most abundant phytoplankton genera identified previously by microscopy, however, were not well represented on the microarray. The lack of sequence data for well-studied species, and the inability to identify organisms associated with functional gene sequences in the environment, still limits our understanding of phytoplankton ecology even in this relatively well-studied system.

KEYWORDS: phytoplankton; microarray; functional gene; North Atlantic; spring bloom

INTRODUCTION

The spring bloom in the North Atlantic Ocean is one of the biggest biogeochemical signals on earth and dominates the annual primary productivity of the region. The bloom develops rapidly in response to changes in heat, light and stratification in the early spring. Ever since Sverdrup's critical depth hypothesis

available online at www.plankt.oxfordjournals.org

(Sverdrup, 1953), the explicit mechanisms that allow the bloom to develop and control its timing have been debated (Henson *et al.*, 2006; Behrenfeld and Boss, 2014). Regardless of the relative importance of mixing, turbulence, grazing, light, etc., the end result is a diatom-dominated bloom in which a few species dominate both the biomass and the vertical flux (e.g. Joint *et al.*, 1993; Weeks *et al.*, 1993; Savidge *et al.*, 1995; Rynearson *et al.*, 2013).

A multiplatform experiment was undertaken in spring 2008 in order to gather high-resolution measurements within a single phytoplankton patch during the spring bloom [Alkire *et al.*, 2012; Mahadevan *et al.*, 2012; the 2008 North Atlantic Bloom (NAB) experiment]. The bloom in April and May of 2008 was initiated up to a month earlier than could have resulted from warming alone, due to eddy-driven stratification that prevented mixing and enabled a patchy, diatom-dominated bloom (Mahadevan *et al.*, 2012). Our sampling occurred during the "main bloom" period (Alkire *et al.*, 2012) in early May at a subset of the stations described by Rynearson *et al.* (2013).

We investigated the composition of the phytoplankton assemblage using key functional genes that encode N and C uptake/assimilation in eukaryotic and prokaryotic phytoplankton. This suite of genes was represented on a microarray, referred to as the phytoarray, which contains 258 archetype probes for genes encoding the enzymes nitrate reductase, nitrate transporters and ribulose bisphosphate carboxylase (RuBisCO). Each probe sequence hybridizes with sequences in the target sample that represents an archetype—any sequence within 87% identity to the probe sequence. Thus, the array is targeted toward known genes, but includes probes derived from environmental sequences as well as those representing species in culture. Because of the archetype approach, i.e. each probe hybridizes with its own perfect match sequence but also with closely related sequences, the array can detect unknown sequences. The relatively large number of probes provides much greater resolution than can be obtained by optical indices or pigment analysis and has advantages over microscopy of analysis time and detection of unknown types. By focusing on functional genes, we may be able to link community composition to environmental adaptation and selection more directly than using identification based on ribosomal RNA gene sequences.

The hybridization behavior of the functional gene microarray design upon which the phytoarray is based has been characterized previously (Taroncher-Oldenburg *et al.* (2003) quantified the behavior of the 70-mer oligoprobe approach; Ward (2008) and Ward *et al.* (2011) described an earlier version of the phytoarray and Ward and Bouskill (2011) described the specifics of the method and validated PCR and whole genome target approaches). Briefly, the archetype probes were selected using an iterative algorithm developed by George Jackson (Bulow *et al.*, 2008) that is designed to select sequences that represent the entire sequence database for a particular gene in a way analogous to that used to identify operational taxonomic units (OTUs; Schloss and Handlesman, 2005). The probe set contains the smallest number of probes that will hybridize with at least one of the sequences in the database (determined by an identity level of 87 ± 3%) (Taroncher-Oldenburg *et al.*, 2003). Each archetype probe will hybridize with sequences >~87% identity, and the strength of the signal increases with increasing identity. Therefore, an abundant sequence with an identity of 87% might produce a signal as intense as that of a much less abundant sequence with 100% sequence identity. For that reason, the array cannot be used to determine absolute abundance. For convenience, we refer to the hybridization results in terms of strength of hybridization signal or intensity or as relative abundance, focusing on the differences between samples for individual probes, rather than assigning too much importance to the absolute signal strength.

We hypothesized that the phytoplankton assemblage would exhibit distinct temporal and spatial (T/S) patterns during the North Atlantic spring bloom that were linked to the physical and chemical environment. The bloom composition is usually evaluated using microscopy, which has identified particular species as major dominants in the assemblage and documented successional changes during the bloom development. Are the same T/S patterns evident at the level of gene sequences? Can we identify different patterns for particular archetypes within the overall bloom assemblage? How do the assemblage patterns described by gene sequences compare to those described by biogeochemical methods? With the phytoarray, we used gene sequences to chart the relationships of phytoplankton types to environmental variables and to other members of the assemblage.

METHOD

Sampling overview

Suspended particulate material was collected during the NAB experiment cruise on the R/V Knorr KN193-03, 2–20 May 2008, YearDay (YD) 125–139 (Table I, Fig. 1) by Tatiana Rynearson. Shipboard samples were collected from near-surface depths in the Iceland Basin in conjunction with a passively drifting, mixed-layer Lagrangian float as described by Rynearson *et al.* (2013). Although the samples are identified by YD in order to make them directly comparable to other reports from the same cruise, the sampling scheme conflates T/S variability because the samples were collected at variable distances

Table I: Station location and sample environmental information

YD	Sample ID	Lat N	Long W	Depth (m)	*T*	*S*	Sigma theta	O_2	PAR	bbp 700	beam_cp	Chl	Phaeo	POC	NO_3	Si
125	YD125	60.85	26.64	4.7	8.91	35.27	27.34	278.81	564.83	0.0022	0.2863	2.08	0.26	112.86	10.84	4.03
126	YD126	60.92	27.00	4.3	8.92	35.28	27.35	290.47	701.89	0.0026	0.3208	2.33	0.64	120.11	*10.00*	*3.00*
127	YD127	60.10	27.41	4.7	8.49	35.23	27.38	299.29	322.32	0.003	0.3074	1.88	0.52	107.83	11.04	2.33
128	YD128	60.10	26.66	4.5	9.03	35.23	27.33	305.67	783.66	0.0039	0.5575	2.34	1.25	151.20	9.3157	1.30
129	YD129	61.07	26.66	4.9	9.21	35.23	27.30	306.99	0.8037	0.0049	0.567	3.46	1.07	211.50	8.5691	0.64
131	YD131A	61.40	26.22	4.3	9.14	35.25	27.29	300.21	186.84	0.0045	0.5272	3.04	0.98	147.15	8.0391	0.50
131	YD131B	61.40	26.22	4.3	9.14	35.25	27.29	300.21	186.84	0.0045	0.5272	3.04	0.98	147.15	8.0391	0.50
132	YD132	61.44	25.98	4	9.21	35.25	27.28	302.79	290.48	0.0047	0.597	3.93	1.07	174.89	9.66	0.15
134	YD134A	61.20	26.12	4.8	9.58	35.26	27.23	310.97	240.61	0.0047	0.6543	2.69	2.51	229.65	7.49	0.44
134	YD134B	61.20	26.12	4.8	9.58	35.26	27.23	310.97	240.61	0.0047	0.6543	2.69	2.51	229.65	7.49	0.44
137	YD137A	61.46	25.95	4.5	9.45	35.26	27.24	299.54	227.61	0.0033	0.3531	0.92	0.26	149.36	8.5298	1.19
137	YD137B	61.46	25.95	4.5	9.45	35.26	27.24	299.54	227.61	0.0033	0.3531	0.92	0.26	149.36	8.5298	1.19
139	YD139	61.23	25.53	3.8	9.49	35.25	27.24	293.37	539.18	0.0028	0.2537	0.89	0.19	101.25	7.6407	0.71

YD, YearDay; *z*, depth, m; *T*, temperature, °C; *S*, salinity, practical salinity unit; θ, sigma theta, potential density, kg m^{-3}; O_2, dissolved oxygen, μM; PAR, photosynthetically active radiation, μmol photon m^{-2} sec^{-1}; bbp 700, particulate backscattering coefficient, m^{-1}; beam_cp, particulate scattering coefficient, m^{-1}; Chl, chlorophyll *a*, μg l^{-1}; Phaeo, phaeopigments, μg l^{-1}; POC, particulate organic carbon, mg m^{-3}; NO_3, NO_2^- + NO_3^-, μM; Si, silicic acid, μM. The values in italics were estimated from the closest cast but were not available for the cast on which the samples were collected.

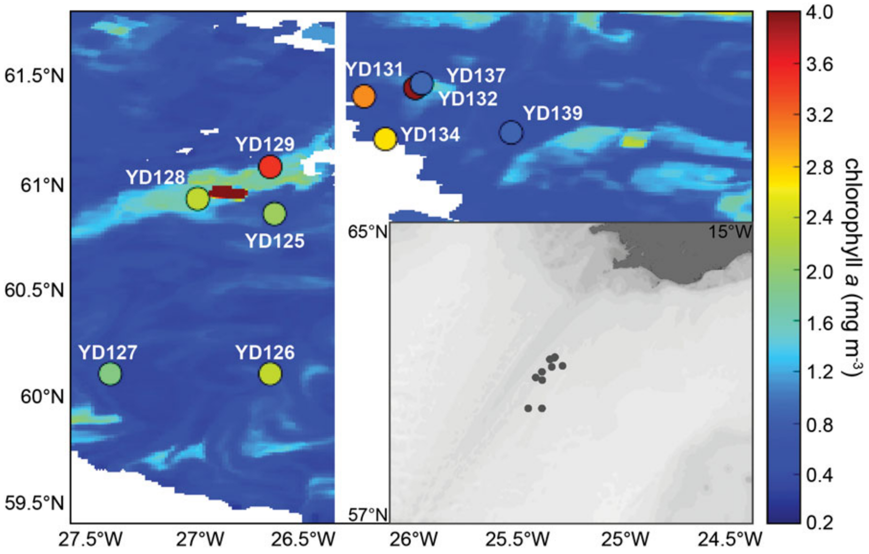

Fig. 1. Composite of two remote sensing chlorophyll *a* concentration images during the Main Bloom (west, 7 May 2008, YD128) and bloom Termination (east, 12 May 2008, YD134) periods (merged MODIS and MERIS CHL1 products at 1 km resolution, from GlobColour, Acri-ST; O'reilly *et al.*, 2000). Colored dots represent the *in situ* surface chlorophyll a concentration on a particular YD (Table 1). Inset map indicates the location of the stations relative to Iceland to the north.

from the center of the bloom. Hydrographic and nutrient data were obtained from BCO DMO (http://data.bco-dmo.org/jg/serv/BCO/NAB08/).

DNA and RNA extraction

Seawater samples (up to 2 L) were filtered onto 0.2-μm pore size Sterivex filters (Millipore, Billerica, MA) using a peristaltic pump and filters were flash frozen in liquid nitrogen and stored at −80°C. Total DNA and RNA was extracted from Sterivex filters using the AllPrep DNA/RNA Mini Kit (Qiagen Sciences, Germantown, MD) with slight modifications to the manufacturer's instructions. The extraction procedures were performed twice on each Sterivex filter in order to maximize the DNA yield and the DNA was stored at −80°C until further processing.

Probe selection and array design

The archetype array approach used in this study has been published previously (Bulow *et al.*, 2008; Ward and Bouskill, 2011). The array used, BC013, contained a total of 258 archetype probes representing genes involved in C and N assimilation in eukaryotic and cyanobacterial phytoplankton. An established algorithm (Bulow *et al.*, 2008) was used to design archetype probes for 12 probe sets (Table 2) corresponding to broad phylogenetic groups within the functional genes of interest, representing sequences from GENBANK at the time of probe design (April 2009). Each 90-mer oligonucleotide probe consisted of a 70-mer archetype sequence (Supplemental Table 1) combined with a 20-mer reference oligo as an internal standard. The algorithm used for probe selection ranks the archetypes according to the number of non-overlapping sequences in the database that should hybridize with that probe. For example, Diatom_rbcL_1 represents the largest number (163) and Diatom_rbcL_2 represents the next largest number (44), of the total diatom *rbcL* sequences (371) used in the probe selection algorithm (Supplemental Table 1). In many of the probe sets, most of the probes were defined by a single sequence at the time of the array design and many did not include any known cultured representatives. The algorithm identifies the minimal number of probes needed to allow hybridization with every sequence in the database while minimizing the number of sequences that might hybridize with more than one probe.

Targets for microarray hybridization were prepared from whole DNA preparations, hybridized in duplicate on a microarray slide and washed as described previously (Ward and Bouskill, 2011). Washed slides were scanned using a laser scanner 4300 (Agilent Technologies, Palo Alto, CA) and analyzed with GenePix Pro 6.0 (Molecular Devices, Sunnyvale, CA).

Data analysis

Quantification of hybridization signals was performed as described previously (Ward and Bouskill, 2011). A normalized fluorescence ratio (FRn) for each archetype was calculated by dividing the fluorescence signal of the archetype by the highest fluorescence signal within the same probe set. The FRn of each archetype from the duplicate arrays was averaged. The relative fluorescence ratio (RFR) of each archetype was calculated as the contribution of FRn of the archetype to the sum of FRn of all archetypes within each probe set. The original array data are available at Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/projects/geo/) at the National Center for Biotechnology Information under GEO Accession No. GSE81262.

Statistical analysis

The array samples are identified by YD and biological replicate. For example, YD131A and YD131B represent two different samples collected from the same Niskin bottle, but analyzed separately in duplicate. Except for the temporal pattern analysis (below), biological replicates are treated as separate samples. The array data were analyzed using the "vegan" package in R (http://www.R-project.org) (Borcard *et al.*, 2011). RFR values were arcsine–square root transformed to normalize the proportional data. Environmental data were square root transformed and then standardized around 0 (decostand in vegan). Highly correlated variables were omitted (Supplemental Table 2). For example, the two scattering measures bbp 700 (particulate backscattering coefficient, m^{-1}) and

Table II: List of probe sets and characteristics

Gene	Phytoplankton target group	Probe set	Number of database sequences	Number of probes	Number of probes per sequence
RuBisCO *(rbcL)*	Chromophytes		**639**	**78**	0.12
		Diatom_rbcL	371	24	0.06
		Hapto_rbcL	133	19	0.14
		NDNH_rbcL	105	25	0.24
		DistantChromo_rbcL	30	11	0.37
	Chlorophytes	Chloro_rbcL	**54**	**16**	0.30
	Cyanobacteria	Cyano_rbcL	**62**	**13**	0.21
Nitrate reductase *(NR, narB)*	Chromophytes		**326**	**62**	0.19
		Diatom_NR	239	45	0.19
		NonDiatom_NR	87	17	0.20
	Chlorophytes	Chloro_NR	**54**	**24**	0.44
	Cyanobacteria	Cyano_narB	**110**	**13**	0.12
Nitrate transporter *(Nrt2, hnat)*	Eukaryotic algae	Euk_Nrt2	**205**	**49**	0.24
	Cyanobacteria	Cyano_hnat	**35**	**4**	0.11

Number of database sequences = number of sequences from public databases that defined each probe based on the algorithm of Bulow *et al.* (2008); the algorithm thus determined the number of probes required to hybridize with all the sequences in the database. The probes per sequences ratio is a measure of the divergence among the available sequences. Diatom_rbcL, diatom *rbcL*; Hapto_rbcL, haptophyte *rbcL*; NDNH_rbcL, non-diatom non-haptophyte chromophyte *rbcL*; DistantChromo_*rbcL*, distant chromophyte *rbcL*; Diatom_NR, diatom *NR*; NonDiatom_NR, non-diatom chromophyte *NR*.

beam_cp (particulate scattering coefficient, m^{-1}) were almost perfectly correlated so only bbp 700 was included in the analysis. Principal components analysis (PCA) was performed using the transformed environmental data and FRn of each archetype.

Investigation of variations in assemblage composition among samples was carried out by clustering the archetype signals using Ward's minimum variance clustering (Bouskill *et al.*, 2011). The number of significant clusters was chosen by optimizing the silhouette score (vegan). For the T/S analysis, FRn data for biological replicate arrays were averaged to produce one value per YD (rather than one value per sample).

RESULTS

Because the analysis of phytoplankton assemblage patterns depends entirely upon the phytoarray, the characteristics of the array and the probes sets are described first (and with additional details in the Supplementary Data). Then the phytoplankton assemblage is described in terms of the most abundant archetypes detected by the phytoarray, and lastly, T/S patterns in the bloom sequence are described. "Probe" refers to the sequence of the oligonucleotide on the microarray. "Archetype" refers to the group of sequences that hybridize to the probe and implies a group of organisms that possess the complementary archetype probe sequence (Box 1).

The phytoarray

The functional gene microarray used in this study is a much expanded version of the original phytoarray (Ward, 2008; Ward *et al.*, 2011) and includes probes for genes encoding three enzymes: RuBisCO (*rbcL*), nitrate reductase (*NR* in eukaryotes, *narB* in cyanobacteria) and high affinity nitrate transporter (*Nrt2* in eukaryotes, *hnat* in cyanobacteria). The 258 archetype probes on the phytoarray are divided into 12 major probe sets, each of which represents functional genes from the major phytoplankton groups (Table II). The sequence database as of April 2009 yielded 35–371 sequences for each functional gene type (in bold in Table II). Application of the probe identification algorithm (Bulow *et al.*, 2008) resulted in the assignment of 4–78 archetype probes for each probe set. The chromophyte probe sets are the largest, and they represent the largest sequence database, although they are still incomplete in important ways (see Discussion) and do not by any means represent the entire chromophyte diversity. Six different Chromophyte probe sets were distinguished, four for *rbcL* and two for *NR*. Chlorophyte and cyanobacterial sequences yielded single probe sets each for *rbcL*, *narB* and *hnat*.

Box 1. **Functional Gene:** A gene that encodes an enzyme involved in some metabolic function, as contrasted to ribosomal genes, which encode a portion of the ribosome. The functional gene sequences upon which the probes were based were mostly obtained from clone libraries derived from PCR amplification, either from cultivated strains or from DNA extracted from environmental samples.

Probe: 70 bp oligonucleotide sequence designed to hybridize with sequences having >87% identity to the probe sequence. The probes were selected using an algorithm (Bulow *et al.* 2008) that identifies the 70-mer region of greatest variability within the aligned sequences, then iteratively identifies the set of sequences that represent all of the aligned sequences with the minimum number of probes and the minimum number of overlaps between sequences represented by individual probes. For example, 639 chromophyte *rbcL* sequences were aligned and the algorithm allows them to be represented by 78 probes.

Probe set: A group of probes identified as hybridizing with targets defined by phylogenetic affiliation. For example, all diatom *rbcL* genes should hybridize with one of the probes in the Diatom *rbcL* probe set. In reality, the probe set is not definitive because it is based on known sequences. Nevertheless, sequences which have >87% identity with the probe will be detected, even if the identity of those sequences are unknown.

Archetype: The group of sequences that hybridize to an individual probe. The probe is identified by the name of a single sequence, but it hybridizes to closely related sequences, and that group of target sequences is referred to as an archetype. The archetype includes an unknown number of individual sequences, which vary in identity with the probe from 87 to 100%. Therefore, the archetype is most likely composed of multiple different sequences, not only the one for which the probe with which they hybridize is named.

The names of the 258 archetype probes and their phylogenetic affiliations are provided in the Supplementary Data (Supplemental Table I and phylogenetic trees in the Supplemental Figures). The identities were reevaluated in August 2014 to determine whether more recent additions to the public databases could be used to identify the unknown environmental sequences. The number of new sequences from cultivated strains was very small, so most of the archetypes remain unidentified except by phylogenetic inference.

The sequence database was largest for *rbcL*, especially due to the work of Paul and others (e.g. Paul *et al.*, 1999; Paul *et al.*, 2000; Wawrik *et al.*, 2003; Wawrik and Paul, 2004) exploring the diversity and biogeography of this gene in the marine environment. Among the eukaryotic algae, the *NR* genes are more divergent than the *rbcL* genes, i.e. more probes are required to represent the entire sequence database for *NR* than for *rbcL* (Table II), but the opposite was true for the cyanobacteria.

Many of the *rbcL* probes hybridize with known *rbcL* sequences. However, the relative lack of divergence in the *rbcL* genes means that a signal from these probes could indicate the presence of a number of different species, i.e. it cannot be used to resolve community composition at the genus or species level. The Diatom_rbcL probe set includes probes with specificities for both centric and pennate diatoms. Other Chromophyte algae, including prymnesiophytes/haptophytes, dinoflagellates and pelagophytes, as well as a number of other cultivated types, are represented in the probe sets Hapto_rbcL, NDNH_rbcL (non-diatom non-haptophyte) and Distant_Chromo_rbcL. Thirty of the Chromophyte *rbcL* sequences were so divergent as to cluster completely separately from the other 609 sequences in the analysis, so they were analyzed separately and they comprise the Distant_Chromo_rbcL probe set. A recent BLAST analysis of these sequences identifies them as cryptophytes, haptophytes, pelagophytes, dinoflagellates and even diatoms, but with sequences that could not be aligned well enough with the other representatives of those clades to include in the probe selection algorithm for the main probe sets.

In contrast to the Diatom_rbcL probes, none of the top 16 Diatom_NR probes, representing 202 out of the 239 sequences in the database at the time of probe development, are predicted to hybridize with any published sequence from a cultivated strain. The clone library studies from which these sequences were derived are not comprehensive in their coverage either geographically or ecologically. This lack of overlap between the clone libraries and the cultured strains, however, does imply that the most abundant *NR* genes present in environments as disparate as the English Channel, Monterey Bay upwelling water, sea grass epiphytes and New Jersey coastal seawater (the main locations of the clone library studies) are not represented in the culture collection. Or, if they are in culture, they have not been sequenced.

Fewer total sequences were available for the Chlorophyte algae, but again, the *rbcL* gene is much less divergent than the *NR* gene. Sequences from known freshwater strains, e.g. *Chlamydomonas*, were omitted from the analysis. As observed for the chromophyte *rbcL* probes, some well-known and widely occurring cultivated strains, *Micromonas pusilla* and *Dunaliella* (*D. salina* and *D. tertiolecta*) are represented in the Chlorophyte *NR* probe set.

There are very few published Chlorophyte *NR* sequences from the environment and only one of the probes is closely related to a known sequences (*Dunaliella* sp.). The dearth of sequences from known marine chlorophytic algae is partly responsible for our inability to provide taxonomic identification for the probe sequences; some of these sequences might represent cultured algae for which sequences are not available. Thus, although the sequences were obtained with primers designed to amplify chlorophyte algal *NR*, this probe set is not well characterized.

The Euk_Nrt2 (eukaryotic high affinity nitrate transporter) probe set is based on sequences obtained with the primers of Song and Ward (2007) and includes mainly diatom and a few other Chromophyte nitrate transporter sequences. The database of sequences from cultivated strains is very limited for this gene, so it is not even possible to make phylogenetic associations for most of the probes.

Most of the cyanobacterial *rbcL* probes have high identity to known cultivated species of *Synechococcus* and *Prochlorococcus* (see Supplementary Data for clade identifications). The cyanobacterial *narB* probes represent several cultivated strains of *Synechococcus* plus many environmental sequences. The *Prochlorococcus narB* gene is not represented on the array for two reasons: when the array was designed, it was not known that *narB* was present in any *Prochlorococcus* genome and, when it was discovered (Martiny *et al.*, 2009), its reported *G/C* ratio is so low that it could not hybridize under the conditions used for the rest of the probes on the array. The four cyanobacterial nitrate transporter probes (Cyano_hnat) all represent known species of *Synechococcus*.

Phytoplankton assemblage composition during the NAB experiment

The phytoarray hybridization data are interpreted in terms of relative signal strength, corresponding qualitatively to relative abundance. These data cannot provide information on the absolute abundance of phytoplankton species represented by individual probes (Ward *et al.*, 2007; Ward, 2008) but at first approximation, the strongest signals (highest FRn or RFR) imply greater relative abundance of the target genes. Then, the most robust interpretation derives from the comparison of relative signal strength between samples.

All of the highest signal *rbcL* archetypes (Table III) have known relatives that should hybridize with the probes (i.e. identity >87%). This was not true for the *NR* probes, where the largest hybridization signals were usually obtained from probes that were rare in the sequence database and often not closely related to known species (Table III). In some cases, even though the closest relative in the database can be assigned to a cultivated strain, the level of identity between the probe sequence and the known relative is so low that it is clear that the probe does not represent that strain, but rather can only be said to represent, e.g. some subset of centric diatoms. For example, the biggest signal for diatom *NR* was Diatom_NR_31, which represented one sequence in the database. This sequence was obtained from a clone library study of epiphytic algae growing on sea grass blades in Tampa Bay, FL (Adhitya *et al.*, 2007). It has no close relatives (i.e. none with an identity level (>87%) close enough to predict hybridization) in the database at the DNA level, but can be identified as a diatom *NR* sequence with 76% identity to *Entomoneis* cf. *alata*. The three diatom *NR* probes that produced the highest hybridization signal (Diatom_NR_31, Non_Diatom_NR_3 and Diatom_NR_35) represented only five total sequences from the 239 *NR* sequences used to derive the probe set, i.e. they were rare in the database but relatively abundant in the samples. Non_Diatom_NR_3 was identified as a diatom by phylogeny only (54% identity with *Amphora* sp.) and its sequence was so distant as to prevent its alignment with the other diatom *NR* sequences, so it was originally designated Non_Diatom. We retain that label because its probe sequence (like the other Non-Diatom_NR probes) represents a different 70-mer region than the Diatom_NR probes.

In 3 of the 12 probe sets, the #1 probe (i.e. the probe that represented the largest number of published sequences) was among the strongest signals in some of the samples (Table III). Chloro_NR_1, with 86% identity to *Ostreococcus lucimarinus*, was the strongest signal in YD125, but was a very minor component in all other samples. For both Cyano_narB and Cyano_hnat probe sets, the #1 probe produced the strongest hybridization signal. These two probes both represent a number of cultivated *Synechococcus* strains and environmental sequences. The #1 probe did not produce the highest Cyanobacterial *rbcL* signal although it was detected in all samples. *Synechococcus* WH8102 sequences are included in all of the #1 cyanobacterial probes (*rbcL*, *narB*, *hnat*). While strong signals from the cyanobacterial probes that are well represented in the database is consistent with cultivated strains being important in the environment, it is also true to the marine cyanobacteria in general are much less diverse than the eukaryotic algae. Thus, it is perhaps not surprising to find more consistency between clone library data and the phytoarray results for the prokaryotes.

The most striking result from the phytoarray analysis of assemblage composition is that YD125 differed from every other sample in every single probe set (Table III and Fig. 4). The other samples, YD126–YD139, were more similar to each other in community composition for all 12 probe sets than any of them were to YD125. The archetypes that yielded the largest signals in each probe set are listed in Table III, and briefly described here.

CHROMOPHYTE ARCHETYPES

Chromophyte *rbcL*

(i) Diatom *rbcL*: the two archetypes with the highest signal after YD125 were Diatom_rbcL_9, representing several sequences from the English Channel (Bhadury and Ward, 2009), with 91% identity with *Gomphonema* sp. a freshwater diatom, and Diatom_rbcL_18, with 100% identity to *Cyclotella* sp., also primarily a freshwater genus. Diatom_rbcL_2, which had 90–97% identity with known *Chaetoceros* sequences, was detected as a major signal on YD125 but as a minor component thereafter.

(ii) Haptophyte *rbcL*: Archetype Hapto_rbcL_7, with 100% identity to *Pleurochrysis haptonemofera*, was a major signal in all samples except YD125, as were archetypes Hapto_rbcL_8 and Hapto_rbcL_15. Hapto_rbcL_8 has highest identity with *Chrysochromulina* sp., and Hapto_rbcL_15 with *Cruciplacolithus* and *Calcidiscus*, and so mostly likely represent members of these genera, although not those in the cultivated database at present.

(iii) NDNH *rbcL*: the three highest signals in this group for all samples except YD125 were NDNH_rbcL_20, with 100% identity to the silicoflagellate *Pseudopedinella elastica*, NDNH_rbcL_7, representing a Chrysophyte with no known relatives with identities high enough for significant hybridization, and NDNH_rbcL_8, which has 100% identity with the dinoflagellate *Karenia mikimotoi*.

(iv) Distant chromophyte *rbcL*: the three highest signals in this group represent a diatom, a dictyophyte and a prymnesiophyte (Table III).

Table III. Highest hybridization signal during the NAB 2008 experiment for each probe set

Gene type		Probe set	Archetype	Closest identity with known relative sequence	Comments
Rubisco	Chromophyte	Diatom_rbcL	Diatom_rbcL_9	*Gomphonema* sp. (91%)	Biggest signal except YD125
			Diatom_rbcL_18	*Cyclotella ocellata* (100%)	2nd biggest signal except YD125
			Diatom_rbcL_23	*Amphiprora alata* (95%)	Biggest signal YD125
		Hapto_rbcL	Hapto_rcbL_7	*Pleurochrysis haptonemofera* (100%)	Major signal except YD125
			Hapto_rcbL_8	*Chrysochromulina spinifera* (86%)	Major signal except YD125
			Hapto_rcbL_15	*Cruciplacolithus neohelis* (90%)	Major signal except YD125
			Hapto_rcbL_17	*Chrysochromulina simplex* (87%)	Biggest signal YD125
		NDNH_rbcL	NDNH_rbcL_20	*Pseudopedinella elastica* (100%)	Major signal except YD125
			NDNH_rbcL_7	Chrysophyte (88%)	Major signal except YD125
			NDNH_rbcL_8	*Karenia mikimotoi* (100%)	Major signal except YD125
			NDNH_rbcL_14	Raphidophyte *Pavlova pinguis* (91%)	Biggest signal YD125
		DistantChromo_rbcL	DistantChromo_rbcL_3	Diatom *Fragilariopsis kerguelensis* (95%)	Major signal except YD125
			DistantChromo_rbcL_9	Dictyophyte *Chattonella verruculosa* (100%)	Major signal except YD125
			DistantChromo_rbcL_10	Prymnesiophyte *Isochrysis galbana* (100%)	Major signal except YD125
	Chlorophyte	Chloro_rbcL	Chloro_rbcL_5	Desmodesmus sp (91%)	Biggest signal except YD125
			Chloro_rbcL_13	*Pyramimonas grossi* (96%)	Major signal YD125
			Chloro_rbcL_15	*Tetraselmis* aff. *maculata* (97%)	Biggest signal YD125 and YD132, 2nd biggest on other days
	Cyanobacterial	Cyano_rbcL	Cyano_rbcL_5	*Synechococcus elongata* (100%)	Biggest signal except YD125
			Cyano_rbcL_7	*Synechococcus rubescens* (100%)	2nd biggest signal except YD125
			Cyano_rbcL_6	Synechococcus sp. CC9902 (100%)	Biggest signal YD129 - 132
Nitrate reductase	Chromophyte	Diatom_NR	Diatom_NR_31	*Entomoneis* cf. *alata* (76% protein level)	Major signal except YD125
			Non_Diatom_NR_3	Distant diatom	Major signal except YD125
			Diatom_NR_35	*Amphora* (100%)	Major signal except YD125
			Diatom_NR_9	Distant diatom	Biggest signal YD125
		NonDiatom_NR	Non_Diatom_NR_9	Chlorarachnion by phylogeny only	Biggest signal YD126-128
			Non_Diatom_NR_5	*Chlorarachnion reptans* CCMP 238 (94%)	Biggest signal YD125
			Non_Diatom_NR_12	Identifiable as NR only at aa level	Biggest signal after YD128
	Chlorophyte	Chloro_NR	Chloro_NR_12	*Heterosigma akashiwo* (91%)	Biggest signal except YD125
			Chloro_NR_6	*Emiliania huxleyi* (84%)	Minor signal except YD125
			Chloro_NR_23	*Fragaria vesca* (76%)	Minor signal except YD125
			Chloro_NR_1	*Ostreococcus lucimarinus* (86%)	Biggest signal YD125
	Cyanobacterial	Cyano_narb	Cyano_narB_1	Synechococcus WH8102 (100%)	Biggest signal except YD125
			Cyano_narB_4	Synechococcus sp. CC9902 (91%)	Biggest signal on YD125
			Cyano_narB_3	Unknown cyanobacteria	Variable, significant on all days
			Cyano_narB_12	Unknown cyanobacteria	Minor signal on YD125, consistent large signal thereafter
Nitrate transporter	Eukaryotic	EukNrt2	EukNrt2_41	*Skeletonema costatum* (75%)	Biggest signal except YD125
			EukNrt2_35	*Cylindrotheca fusiformis* (73%)	2nd biggest signal except YD125
			EukNrt2_14	Nothing	Biggest signal on YD125
	Cyanobacterial	Cyano_hnat	Cyano_hnat_1	Synechococcus sp. CC9902 (98%)	Major signal on YD125
			Cyano_hnat_2	Synechococcus 8103	Major signal except YD125
			Cyano_hnat_3	Synechococcus sp. RCC307	Major signal except YD125

Percentage values between brackets represent the sequence match between public database sequence and the 70-mer probe. Number of database sequences = number of sequences in the database that defined this probe based on the algorithm of Bulow *et al.* (2008). The archetype probe names are ranked in order of most representation in the sequence database (e.g. the highest signal Diatom_rbcL probe, Diatom_rbcl_9, was the ninth most abundant archetype in the database).

Chromophyte *NR*

(i) Diatom *NR*: Diatom_NR_31 was the largest signal on YD126 and it remained a significant signal for the rest of the experiment. This archetype, representing one uncultivated sequence from the Tampa Bay epiphyte study (Adhitya *et al.*, 2007), has no known close relatives in the database but can be identified as a diatom at the protein level of sequence identity. The second strongest signal, probe NonDiatom_NR_3, was obtained from a clone library from Monterey Bay, CA. The third strongest signal was from the Amphora probe, Diatom_NR_35, representing a cultivated strain.

(ii) Non-Diatom *NR*: NonDiatom_NR_9, representing two sequences from Monterey Bay identified as related to Chlorarachnion (Badhury and Ward, 2009), was the major group for YD126–128. After YD128, archetype NonDiatom_NR_12 became the largest signal. NonDiatom_NR_12 is identifiable as a eukaryotic *NR* sequence only at the protein level, so its phylogenetic affiliation is unknown.

CHLOROPHYTE ARCHETYPES

Chloro_rbcL

Chloro_rbcL_5 was the dominant archetype in all samples except YD125 and YD132. This sequence is a clone from the Gulf of Mexico, which was identified by phylogeny as a chlorophyte, but cannot be identified with more resolution. It has ~90% identity with a large number of chlorophyte cultures. Two other Chloro_rbcL archetypes were important in YD125 but less so in other samples: Chloro_rbcL_13 is another clone sequence from the Gulf of Mexico, but it has 100% identity with many other recent environmental clones from other environments. It is a prasinophyte at 96% identity with various members of the genus *Pyramimonas*. Chloro_rbcL_15 is another clone from the Gulf of Mexico, this one 97% identity with *Tetraselmis* aff. *maculata* and many environmental sequences.

Chloro_NR

The biggest signal across all samples except YD125 was Chloro_NR_12, which has 91% identity with *Heterosigma akashiwo*, a raphidophyte (Chattonellaceae). Archetype Chloro_NR_6 is approximately equally represented in all samples except YD125. It has no close relatives in the DNA database (81% identity with *Hordeum vulgare*, common barley) but is 100% identical to *Emiliania huxleyi*, a chromophyte alga, at the protein level (84% identity at the DNA level). Chloro_NR_23 was the third strongest signal in all but YD125. It is so distant as to find no close relatives at the DNA level, but it is closest to alpine strawberry at the protein level.

Euk_Nrt2

Euk_Nrt2_41 was the biggest signal, and Euk_Nrt2_14 the second biggest, in all samples except YD125. Both of these are distantly related to diatoms. GS52_E3 was the biggest signal in YD125. It has no close relatives in the database but can be shown to be an *Nrt2* gene at the amino acid level.

CYANOBACTERIAL ARCHETYPES

Cyanobacterial archetypes representing commonly cultured strains of *Synechococcus* were well represented in the hybridization signal in all samples, although as for the eukaryotes, the assemblage composition was quite different between YD125 and all other samples. *Prochlorococcus* was always a minor signal compared to *Synechococcus* archetypes, which is consistent with the former's subtropical distribution.

Cyanobacterial *rbcL*

Cyano_rbcL_5, Cyano_rbcL_7 and Cyano_rbcL_6 were the three biggest signals in all samples except YD125. These all represent different *Synechococcus* strains. The four archetypes representing *Prochlorococcus* were detected but at much lower levels.

Cyanobacterial *narB*

Cyano_narB_1 was the dominant archetype in all except YD125. Cyano_narB_12 was an approximately equal portion of the signal in all samples except YD125. Cyano_narB_3 and Cyano_narB_13 both represent unknown cyanobacteria *narB* and were major signals in most samples. Cyano_narB_4, related to *Synechococcus*, was important in YD125 and also YD129 and YD131.

Cyanobacterial *hnat*

All four Cyanobacterial *hnat* sequences represent *Synechococcus* strains and two of them, Cyano_hnat_2 and Cyano_hnat_3, dominated the signal in all samples except YD125, when Cyano_hnat_1 was the biggest signal.

ENVIRONMENTAL SETTING

The development and evolution of the bloom were described by Alkire *et al.* (2012), who identified six periods based on physical and biogeochemical data collected by a suite of floats and gliders. The samples analyzed here were collected during the main bloom and termination period (YD124–134) and the eddy period (YD135–141), the latter so named because the floats encountered a small anticyclonic eddy at that point. The array samples cannot be considered a simple time series, however, because the ship zigzagged across the chlorophyll features that were tracked by the floats and gliders (Fig. 1 and Figs 1 and 5 in Alkire *et al.*, 2012). Thus, the main source of variability is small-scale spatial variability, rather than a linear development; these are not samples from the same patch of water over time. Nonetheless, the chemical and physical data from the depths sampled for the arrays do show a temporal progression that is consistent with the general bloom conditions, including an increase in stratification and a gradual decline in nitrate and silicate concentrations. Silicate was depleted (<0.5 μM) by YD132, while nitrate remained abundant (>7.0 μM), and chlorophyll *a* was highest (>3.0 μg l^{-1}) between YD129 and YD134 (Fig. 2).

PCA clearly identified the sequential change in the environmental setting and integrated biological data (i.e. pigments) for the samples included in the microarray analysis. The samples form a progression on the plot in a clockwise spiral (Fig. S1). Greatest mixed-layer depth (MLD) was associated with the early days of the sequence (YD125, YD126, YD127), while stability (N2) was greatest at the end (YD137, YD139). Highest nutrient conditions (nitrate and silicate) occurred early in the sequence, although neither nutrient was exhausted by the end of the study. Chlorophyll *a* concentration was highest near the middle of the sequence (YD129, YD131, YD132). This analysis simply describes the progression of the oceanic conditions during the 2 weeks of the sampling.

PHYTOPLANKTON ASSEMBLAGE VARIATION WITH TIME/SPACE

Unlike the environmental data, PCA of the phytoarray data did not show a simple temporal progression (Fig. S2) among samples. The relative contribution of different groups varied with time, but not in a unidirectional manner. The community composition of YD125 was most different from all other samples, as already noted (Table III). Simple community diversity analyses (Shannon, Simpson, Pielou's evenness; data not shown) varied minimally and did not reveal consistent patterns with time.

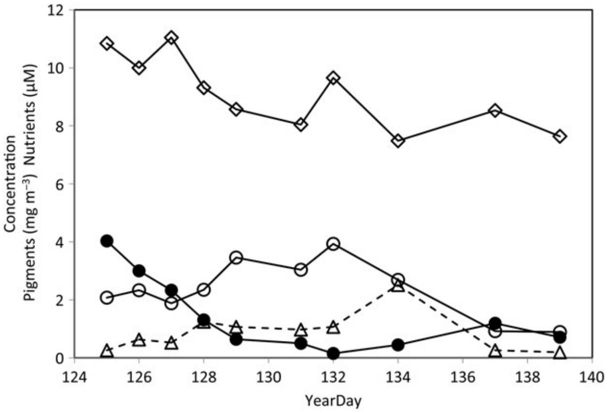

Fig. 2. Nutrient and pigment concentrations for the array samples by YD. [NO_3^-], open diamonds; [Si], closed circles; chlorophyll *a*, open circles; phaeopigments, open triangles.

To assess what environmental variables might explain the observed patterns in assemblage composition and relative importance of individual probe sets, a redundancy analysis (RDA) was performed by combining assemblage composition (in terms of RFR) and environmental data. The FRn values for each archetype set were summed and used to calculate RFR for each set, so that, e.g. Diatom_NR is considered as a fraction of the total signal, i.e. relative abundance of the sum of FRn for all 12 archetype subsets. Cyano_narB, Cyano_rbcL, Chloro_NR and Euk_Nrt2 (which is consisted mainly of chlorophyte archetypes) were associated with YD125 in the environmental sequence. The YD125 assemblage was positively associated with higher MLD, higher nutrient and chlorophyll *a* concentrations and photosynthetically active radiation levels and negatively with N2 (Fig. 3, S2). Diatoms (Diatom_NR and Diatom_rbcL) and NonDiatom_NR were all associated with higher N2 and lower nutrient concentrations, found in the later days of the study (Fig. 3). Assemblages represented by YD137 and YD139 correlated with the diatom probes and N2 (Fig. S2). DistantChromo_rbcL and Cyano_rbcL, the two archetype sets whose average distributional patterns were not correlated with any of the others, were not related to stability (N2) but were negatively correlated with MLD.

Because the RFR data are relative and across the entire assemblage always sum to 1.0, they cannot be used to evaluate absolute abundance of individual groups. FRn is also a relative abundance measure but changes in FRn do reflect absolute changes in signal strength independent of other members of the assemblage. The value of FRn ranges from 0 to 1.0, with 1.0 being the maximum hybridization signal for that probe set. Higher average FRn thus means that more of the individual archetypes had higher signals, independent of the total number of archetypes within the group. Here

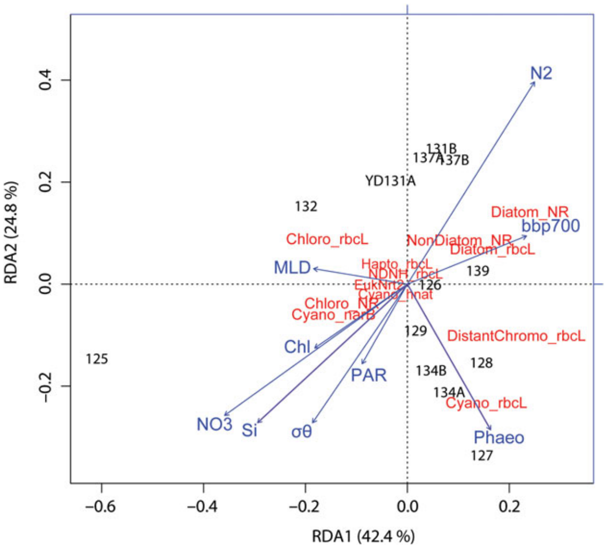

Fig. 3. RDA triplot of 13 array samples (identified by YD) with environmental parameters (defined in Table I) as explanatory variables and archetype probe sets (RFR; abbreviations in Table II) as response variables. In this scaling, the angles between all variables approximate their linear correlations. The right angle projections of the object points (YD samples) onto the vectors reflect qualitatively their correlations with those vectors. For example, YD129 and both YD134 samples had highest values for Phaeo, and YD125 was associated with highest Si and NO3 values.

we use FRn to evaluate T/S patterns among the archetype groups.

Archetype average patterns

T/S variation is first considered on the basis of the major probe sets. Average FRn for each archetype set (e.g. Diatom_rbcL or Diatom_NR) was computed in order to compare patterns between sets.

The two sets of Chromo_NR archetypes (Diatom_NR and NonDiatom_NR) co-varied (Fig. 4) and increased on average between YD125 and YD132, with a second peak on YD137. All four *NR* archetype groups had highest average FRn on YD132. Most of the six *rbcL* archetype sets also had a maximum average FRn on YD132 (Fig. 5). This peak coincides with the maximum chlorophyll *a* concentration (Fig. 2).

Average FRn for Diatom_rbcL was significantly positively correlated ($P < 0.01$) with average FRn for several other archetype sets: Hapto_rbcL, Chloro_rbcL, Diatom_NR, Chloro_NR, Cyano_narB, Euk_Nrt2 and Cyano_hnat (Supplemental Table 3). The two archetype groups representing diatoms (Diatom_rbcL and Diatom_NR) and the two groups representing chlorophytes (Chloro_rbcL and Chloro_NR) were correlated with each other and with the Euk_Nrt2 archetypes ($P < 0.01$). Of the three archetype sets representing

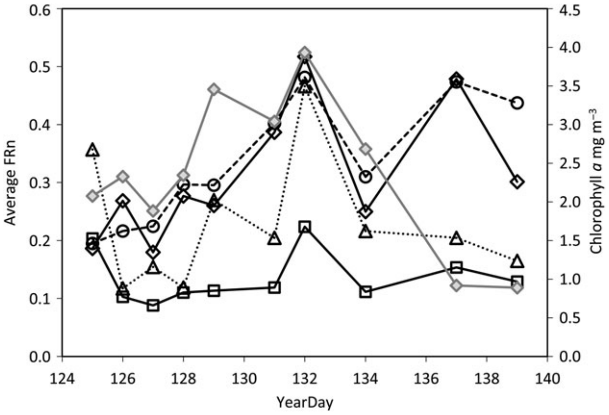

Fig. 4. Time course of average FRn values for the *NR* and *narB* archetype sets and chlorophyll *a*. FRn average was computed for each archetype set, and FRn for biological replicates were averaged. Diatom_NR, open diamonds; NonDiatom_NR, open circles; Chloro_NR, open squares; Cyano_narB, open triangles; Chlorophyll *a*, gray diamonds.

cyanobacteria, Cyano_narB and Cyano_hnat, were correlated but neither was correlated with Cyano_rbcL. None of the individual archetype sets FRn's were significantly correlated with chlorophyll *a* across the samples ($P > 0.05$).

T/S patterns within archetype groups

Variability in the community composition results from both T/S variability in the environment. So although the samples are identified by YD, we will refer to patterns as T/S patterns, acknowledging that we cannot separate the two effects. For the larger probe sets, distinct T/SPs among different archetypes could be detected using cluster analysis. The 23 Diatom_rbcL archetypes clustered into 2 T/S patterns, 11 archetypes in T/SP-1 and 10 in T/SP-2. The major feature of both T/SPs was the maximum in FRn on YD132 (Fig. 6). The two archetypes with the largest hybridization signal did not cluster with the rest of the Diatom_rbcL archetypes but had quite different T/S patterns. Diatom_rbcL_9 had a very strong signal in every sample except YD125 and Diatom_rbcL_18 had its strongest signal in samples where all other Diatom_rbcL archetypes were minimal (Fig. 6).

Hapto_rbcL clustered into five T/SPs, with a consistent peak in YD132 in T/SP-1, T/SP-2 and T/SP-3 (Fig. S3). The basis of the three separate clusters is the scale of the signals: T/SP-1 had the smallest signal, T/SP-2 and T/SP-3 each larger. Of the three major archetypes in Hapto_rbcL (Table III), one was the largest component of T/SP-1 (Hapto_rbcL_8) and the other two clustered separately as T/SP-4, on the basis of the magnitude of their signals rather than similarity

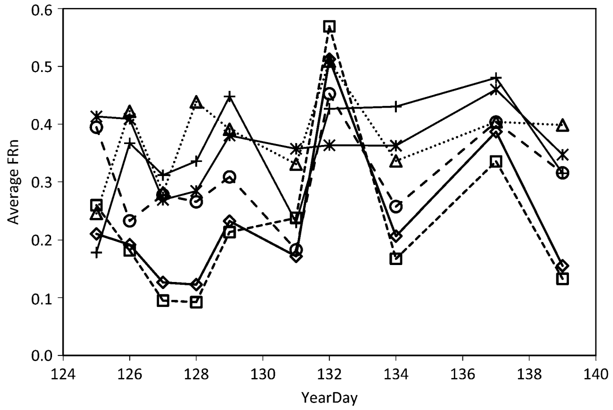

Fig. 5. Time course of average FRn values for the *rbcL* archetype sets. FRn average was computed for each archetype set, and FRn for biological replicates were averaged. Diatom_rbcL, open diamonds; NDNH_rbcL, plus sign; Hapto_rbcL, open circles; DistantChromo_rbcL, star; Chloro_rbcL, open squares; Cyano_rbcL, open triangles.

of their patterns (Fig. S3). The last T/SP had only one member and had a negligible signal in most samples.

NDNH_rbcL clustered into three T/SP's, which were all quite distinct from each other. The three archetypes with the highest FRn signals all clustered together as T/SP-1, but separately from all other NDNH_rbcL archetypes (Fig. S4).

Diatom_NR archetypes clustered into four T/S patterns. T/SP-1 and T/SP-2 contained 15 and 22 of the 45 Diatom_NR archetypes and had similar patterns, with maxima at YD132. T/SP-3 contained 5 archetypes with some of the highest FRn and had a minimum in YD132 (Fig. 7). T/SP-4 represented 3 archetypes, only one of which had significant signal (not shown).

Euk_Nrt2 archetypes clustered into six T/S patterns. For T/SP-1 through T/SP-4, a maximum in YD132 was a common feature (Fig. S5). The two archetypes with the highest overall signal clustered separately as T/SP-6 and did not have a maximum in YD132.

The smaller number of cyanobacterial probes precluded investigation of T/S patterns within subsets of those three probe sets. All three cyanobacterial probe sets had maximal signal in YD132 (Figs 4 and 5; Cyano_hnat not shown).

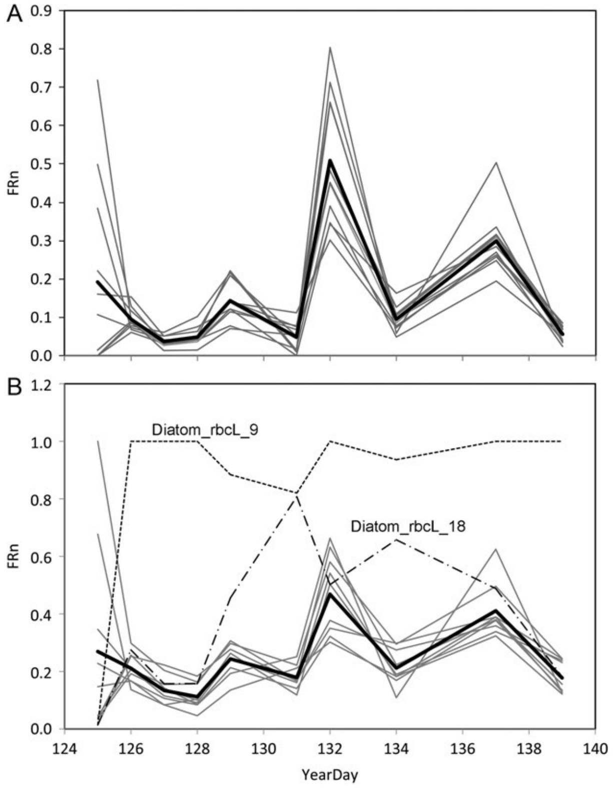

Fig. 6. T/S patterns for the individual archetypes within the Diatom_rcbL archetype sets. Individual archetypes are shown as thin black lines and the centroid of the group is shown as the thick red line. (A) Diatom_rbcL T/SP-1. Individual archetypes are shown as thin black lines and the centroid of the group is shown as the thick red line. (B) Diatom_rbcL T/SP-2. Individual archetypes are shown as thin black lines and the centroid of the group is shown as the thick red line. The dotted and dashed lines represent Diatom_rbcL_9 and Diatom_rbcL_18, which did not cluster with either of the main Diatom_rbcl T/SPs.

DISCUSSION

Assemblage composition and temporal patterns

The environmental data showed a clear spatiotemporal evolution from conditions of deep MLD, low water column stability and high nutrient concentrations to conditions of shallower MLD, stronger water column stability and lower nutrient concentrations. The phytoplankton assemblage described by the phytoarray did not undergo a simple succession, but the relative abundance of diatoms was correlated with greater stability and lower nutrient concentration associated with later stages of the sampling period. Although the assemblage was quite different in the first sample, YD125, all the archetype groups had peak signal strength in YD132, which coincided with highest chlorophyll *a* concentration. The lack of significant variation in diversity and evenness in the phytoplankton assemblage (computed from FRn) over time suggests, however, that the dramatic successional changes associated with bloom formation occurred prior to this sampling window, and big changes associated with the demise of the bloom had not yet manifested. The assemblage analysis distinguished important components among the chromophytic and chlorophytic algae represented on the array, many of which could not be assigned a phylogenetic identity from published sequence databases.

The major features of the T/S patterns, both from the average FRn of the major archetype sets (Figs 4 and 6) and from the separate T/SP analysis of archetype subsets (Figs 6 and 7 and S3-S5), were the maximum in FRn signal strength on YD132 and the difference in composition between YD125 and all other samples. Although the T/S patterns were not correlated with chlorophyll *a* concentration, the maximum in chlorophyll *a* coincided with the maximum in FRn signal strength in YD132 (Fig. 2; Rynearson *et al.*, 2013). YD134 and later are referred to as the bloom termination period (Alkire *et al.*, 2012), which was accompanied by a major sedimentation event, consisted heavily of diatom spores (Rynearson *et al.*, 2013). POC and O_2 concentration integrated over the MLD were maximal around YD132–YD135 (Alkire *et al.*, 2012), even though discrete chlorophyll *a* concentration and chlorophyll *a* concentration integrated over the top 100 m both decreased during this period (Mahadevan *et al.*, 2012).

The microarray data are referred to here as a time course, and the samples were collected sequentially around a Lagrangian float, which did accurately follow a coherent patch and obtained sequential samples of the same patch (Alkire *et al.*, 2012). The microarray samples, however, were collected as discrete shipboard samples and the cruise track passed across and through the patch (Fig. 1), so the array data reflect both the spatial patchiness of the bloom and a time sequence. The fact that variations in simple diversity measures were minimal probably results from the conflation of T/S variability, and the fact that all of the samples except YD125 were collected from the main and termination periods of the bloom. YD125 fell between the early and main bloom periods defined by Alkire *et al.* (2012), which may explain its different community composition in comparison to the other array samples.

The temporal evolution of the phytoplankton assemblage was evaluated by Cetinić *et al.* (2015) using their newly developed optical community index. The index (ratio of chlorophyll *a* fluorescence to backscattering; Chl F/b_{bp}) was measured by instruments on gliders and floats and provided very high temporal resolution. Samples throughout the course of the entire bloom (YD120–YD145) fell into three groups distinguished by the index. The main determinant of the index was percent diatom biomass, due to taxa-specific differences in the chlorophyll *a* to carbon ratio. Consistently high values for the index were recorded for YD120–135, covering most of the period for which microarray data were obtained. The highest values of the optical index were detected in samples YD132–136, although YD131–132 corresponded to a minimum in the diatom contribution to biomass at 10 m assessed by microscopy (Rynearson *et al.*, 2013). YD133 was not included in the array samples, but the FRn peaks in YD132 in several of the T/SPs described by the array data (e.g. Diatom_rbcL T/S P-1, Diatom_NR T/SP-1, 2; Hapto-rbcL T/SP-1, 2; EukNrt2 T/SP-2, 3, 4) are consistent with a shift in the community thereafter, as indicated by decreasing optical density index after YD134 and the

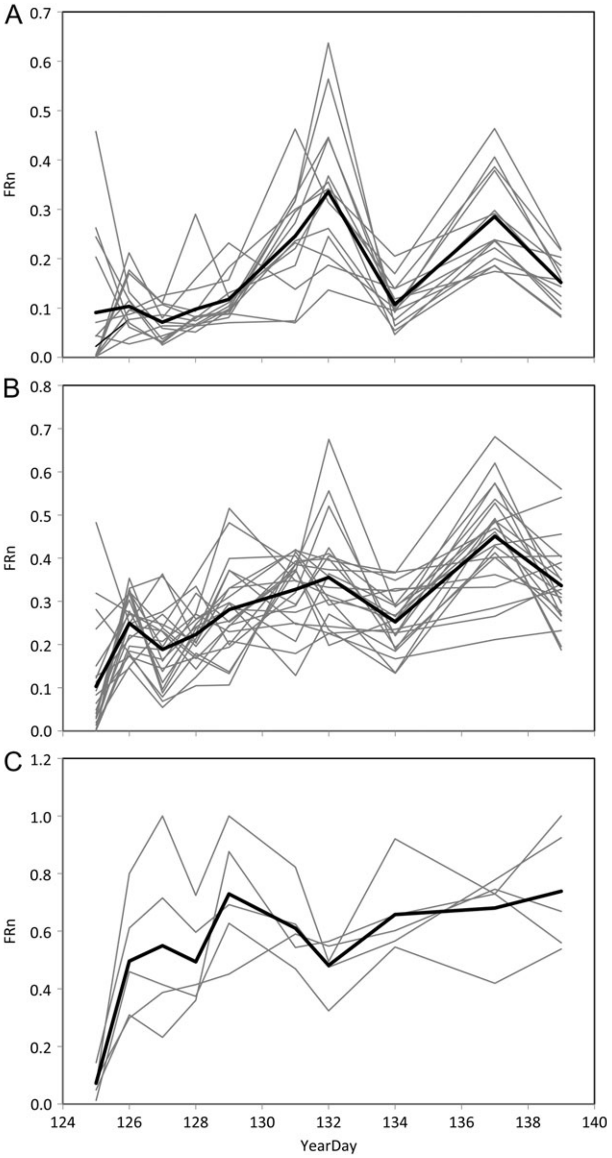

Fig. 7. T/S patterns for the individual archetypes within the Diatom_NR archetype set. Individual archetypes are shown as thin black lines and the centroid of the group is shown as the thick red line. (**A**) Diatom_NR T/SP-1. (**B**) Diatom_NR T/SP-2. (**C**) Diatom_NR T/SP-3.

commencement of a major sinking event at that time (Cetinic *et al.*, 2015). The second peak in FRn coincides with the eddy period (YD135–141).

In all of the archetype sets for which multiple T/SPs were discerned (Figs 6 and 7 and S3-S5), the few archetypes with the strongest signals had T/SPs that were distinct from most of the other archetypes. Thus, most of the assemblage varied together, but the major components behaved differently, suggesting that their dominance is due to their differential ability to exploit the conditions or escape grazing. Such dominance is expected for the main period of a bloom.

The view of the spring bloom from the microarray

Because of the importance of nitrate in new production, and its use by eukaryotic and possibly cyanobacterial phytoplankton in the spring bloom in the North Atlantic, genes involved in nitrate assimilation (*NR* for eukaryotes and *narB* for cyanobacteria) and transport (*Nrt2* for eukaryotes and *hnat* in cyanobacteria), as well as genes for carbon assimilation (*rbcL*) were the focus of the phytoarray. Because most of these sequences were derived from clone libraries (a very small number of them came from genome sequences), it is not possible to link the different probe sets. For example, an *rbcL* sequence may be identified by phylogenetic analysis as highly related to *Thalassiosira weissflogii*, and an *NR* sequence may also be identified as highly related to *T. weissflogii*, but that does not mean that the two unknown sequences derived from the same kind of organism. They are both likely derived from centric diatoms, but not necessarily the same one. Phylogenetic inference is further constrained by the use of 70-mer oligonucleotide probes. The probe region was chosen to maximize resolution/discrimination between similar sequences. Thus, the phylogeny determined from the 70-mer sequences sometimes differs slightly from phylogenetic relationships determined from analysis of longer more complete sequences of the same genes.

Among the *rbcL* probes all of the strongest signals in Table 3 were associated with probes that had identities high enough to hybridize with the most closely related known relatives. For *NR*, however, some of the Chlorophyte and all of the Cyanobacterial *NR* probes had high identity with known species, but only two of the Chromophyte *NR* probes should hybridize with known cultures. Therefore, it is possible to identify important archetypes in the environment from each probe set, but it is often not possible to assign them to any known genus or species. For eukaryotic *NR* and *Nrt2*, even those sequences that were abundant in clone libraries from previous studies are not closely related to any previously sequenced species and cannot be identified beyond class or even order.

Using the probe sequences to identify archetypes are also complicated by different degrees of divergence among the genes: for example, *rbcL* is much more highly conserved than *NR*. So 100% identity between probe and a cultivated strain does not necessarily mean the probe represents that species for *rbcL* (many different species have 100% identity with the probe fragment). However, because *NR* has much more divergence, hybridization with an *NR* probe does restrict the target to a smaller phylogenetic range.

The dependence on clone libraries imposes limitations resulting from the specificity of the PCR primers used to obtain the clones. The public databases at the time of this analysis contained essentially no sequences for *NR*, *Nrt2* or *rbcL* from the major upwelling and spring bloom species, especially members of the diatom genus *Chaetoceros*. The *NR* primers used to develop the clone libraries do not amplify the *Chaetoceros* species tested so far (unpublished data). So it is likely that some of the most abundant species in the world, and in this bloom in particular, are not represented on the array because they are quite different from known sequences, beyond the reach of commonly used PCR primers. Obviously, therefore, our discussion of the community composition is limited to those sequences represented on the array, and this limits the direct comparison between the array results and those reported previously for the 2008 NAB experiment.

The dominant diatoms in the surface waters (detected by microscopy; Rynearson *et al.*, 2013) were *Chaetoceros laciniosus*, *Chaetoceros decipiens* and *Chaetoceros compressus*. No *NR* sequences are available for these species. The *rbcL* sequence for *C. compressus* is now available, but is quite distinct and is not represented on the array. *Ccompressus* aff. *diadema* was not abundant in surface waters, but was the major contribution to the vertical flux that occurred around YD135. No functional gene sequence data are available for *C.* aff. *diadema* at this time. It is possible that some of the unidentified *rbcL* and *NR* sequences on the array do represent *Chaetoceros* genes, but it is also clear that some of the major components of the assemblage were not evaluated in the array data. It should be a high priority to obtain sequence information for species that are documented as numerically important or significant contributors to photosynthetic biomass in the ocean in order to facilitate study of their ecology using high-throughput molecular methods.

CONCLUSIONS

The assemblage of eukaryotic phytoplankton types present throughout the bloom was diverse at the genetic level; most of the 258 archetypes were detected at some level in some samples. The *NR* and *Nrt2* archetypes with the highest hybridization signals were NOT associated with or closely related to sequences from known/cultivated eukaryotic phytoplankton, indicating our lack of knowledge about important phytoplankton species. The strongest *rbcL* signals were due to probes with high identities with known strains, but the relatively low divergence of the *rbcL* gene limits its phylogenetic discrimination. Most of the archetypes exhibited consistent T/S patterns, suggesting that even though these archetype groups represent different individual "species", the phylogenetically defined groups represented ecologically similar relationships. The few archetypes with the largest signals usually exhibited T/S patterns distinct from the rest of the archetypes, suggesting that just a few species dominate the system against a background of high diversity. The mismatch among (i) cultivated species, (ii) species identified by physical/observational methods (e.g. microscopy) as abundant in the environment, (iii) the sequence database (which does not include all cultivated strains) and (iv) sequences (without cultivated close relatives) that are identified as abundant in the environment indicates our ignorance of the ecology of marine phytoplankton. These archetype sequences provide a window into the links between organisms carrying the functional gene and environmental conditions but also point out a huge gap in our knowledge of important eukaryotic phytoplankton even in a system as intensely studied as the North Atlantic spring bloom.

SUPPLEMENTARY DATA

Supplementary data can be found online at http://plankt.oxfordjournals.org

ACKNOWLEDGEMENTS

We are grateful to Mary Jane Perry for the opportunity to participate in this study. We thank Tatiana Rynearson for collecting the samples that we analyzed.

FUNDING

Award NOAA NA11OAR4310058 form the National Oceanic and Atmospheric Administration to BBW.

REFERENCES

Adhitya, A., Thomas, F. I. and Ward, B. B. (2007) Diversity of assimilatory nitrate reductase genes from plankton and epiphytes associated with a seagrass bed. *Microb. Ecol.*, **54**, 587–597.

Alkire, M. B., D'asaro, E., Lee, C., Perry, M. J., Gray, A., Cetinic, I., Briggs, N., Rehm, E. *et al.* (2012) Estimates of net community production and export using high-resolution, lagrangian measurements of O_2, NO_3^-, and POC through the evolution of a spring diatom bloom in the North Atlantic. *Deep Sea Res. Part I: Oceanogr. Res. Pap.*, **64**, 157–174.

Behrenfeld, M. J. and Boss, E. S. (2014) Resurrecting the ecological underpinnings of ocean plankton blooms. *Ann. Rev. Mar. Sci.*, **6**, 167–U208.

Bhadury, P. and Ward, B. B. (2009) Molecular diversity of marine phytoplankton communities based on key functional genes. *J. Phycol.*, **45**, 1335–1347.

Borcard, D., Gilet, F. and Legendre, P. (2011) *Numerical Ecology with R*. Springer, New York.

Bouskill, N. J., Eveillard, D., O'mullan, G., Jackson, G. A. and Ward, B. B. (2011) Seasonal and annual reoccurrence in betaproteobacterial ammonia-oxidizing bacterial population structure. *Environ. Microbiol.*, **13**, 872–886.

Bulow, S. E., Francis, C. A., Jackson, G. A. and Ward, B. B. (2008) Sediment denitrifier community composition and nirS gene expression investigated with functional gene microarrays. *Environ. Microbiol.*, **10**, 3057–3069.

Cetinic, I., Perry, M. J., D'asaro, E., Briggs, N., Poulton, N., Sieracki, M. E. and Lee, C. M. (2015) A simple optical index shows spatial and temporal heterogeneity in phytoplankton community composition during the 2008 North Atlantic Bloom Experiment. *Biogeosciences*, **12**, 2179–2194.

Henson, S. A., Robinson, I., Allen, J. T. and Waniek, J. J. (2006) Effect of meteorological conditions on interannual variability in timing and magnitude of the spring bloom in the Irminger Basin, North Atlantic. *Deep Sea Res. Part I: Oceanogr. Res. Pap.*, **53**, 1601–1615.

Joint, I., Pomroy, A., Savidge, G. and Boyd, P. (1993) Size-fractionated primary productivity in the Northeast Atlantic in May–July 1989. *Deep Sea Res. Part II: Top. Stud. Oceanogr.*, **40**, 423–440.

Mahadevan, A., D'asaro, E., Lee, C. and Perry, M. J. (2012) Eddy-driven stratification initiates North Atlantic spring phytoplankton blooms. *Science*, **337**, 54–58.

Martiny, A. C., Kathuria, S. and Berube, P. M. (2009) Widespread metabolic potential for nitrite and nitrate assimilation among Prochlorococcus ecotypes. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 10787–10792.

O'Reilly, J. E., Maritorena, S., O'Brien, M. C., Siegel, D. A., Toole, D., Menzies, D., Smith, R. S., Mueller, J. L. *et al* (2000) SeaWiFS postlaunch calibration and validation analyses, part 3. In: Hooker, S. B. and Firestone, E. R. (eds), *NASA Technical Memorandum 2000-206892*. NASA Goddard Space Flight Center, pp. 49.

Paul, J. H., Alfreider, A. and Wawrik, B. (2000) Micro- and macrodiversity in rbcL sequences in ambient phytoplankton populations from the southeastern Gulf of Mexico. *Mar. Ecol. Prog. Ser.*, **198**, 9–18.

Paul, J. H., Pichard, S. L., Kang, J. B., Watson, G. M. F. and Tabita, F. R. (1999) Evidence for a clade-specific temporal and spatial separation in ribulose bisphosphate carboxylase gene expression in phytoplankton populations off Cape Hatteras and Bermuda. *Limnol. Oceanogr.*, **44**, 12–23.

Rynearson, T. A., Richardson, K., Lampitt, R. S., Sieracki, M. E., Poulton, A. J., Lyngsgaard, M. M. and Perry, M. J. (2013) Major contribution of diatom resting spores to vertical flux in the sub-polar North Atlantic. *Deep Sea Res. Part I: Oceanogr. Res. Pap.*, **82**, 60–71.

Savidge, G., Boyd, P., Pomroy, A., Harbour, D. and Joint, I. (1995) Phytoplankton production and biomass estimates in the Northeast Atlantic-Ocean, May to June 1990. *Deep Sea Res. Part I: Oceanogr. Res. Pap.*, **42**, 599–617.

Schloss, P. D. and Handlesman, J. (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.*, **71**, 1501–1506.

Song, B. K. and Ward, B. B. (2007) Molecular cloning and characterization of high affinity nitrate transporters in marine phytoplankton. *J. Phycol.*, **43**, 542–552.

Sverdrup, H. U. (1953) On the conditions for the vernal blooming of phytoplankton. *J. du Cons.*, **16**, 287–295.

Taroncher-Oldenburg, G., Griner, E., Francis, C. A. and Ward, B. B. (2003) Oligonucleotide microarray for the study of functional gene diversity of the nitrogen cycle in the environment. *Appl. Environ. Microbiol.*, **69**, 1159–1171.

Ward, B. B. (2008) Phytoplankton community composition and gene expression of functional genes involved in carbon and nitrogen assimilation. *J. Phycol.*, **44**, 1490–1503.

Ward, B. B. and Bouskill, N. J. (2011) The utility of functional gene arrays for assessing community composition, relative abundance, and distribution of ammonia-oxidizing bacteria and archaea *Methods Enzymol.*, **Vol. 496**. pp. 373–396.

Ward, B. B., Eveillard, D., Kirshtein, J. D., Nelson, J. D., Voytek, M. A. and Jackson, G. A. (2007) Ammonia-oxidizing bacterial community composition in estuarine and oceanic environments assessed using a functional gene microarray. *Environ. Microbiol.*, **9**, 2522–2538.

Ward, B. B., Rees, A. P., Somerfield, P. J. and Joint, I. (2011) Linking phytoplankton community composition to seasonal changes in f ratio. *ISME. J.*, **5**, 1759–1770.

Wawrik, B. and Paul, J. H. (2004) Phytoplankton community structure and productivity along the axis of the Mississippi River plume in oligotrophic Gulf of Mexico waters. *Aquat. Microb. Ecol.*, **35**, 185–196.

Wawrik, B., Paul, J. H., Campbell, L., Griffin, D., Houchin, L., Fuentes-Ortega, A. and Muller-Karger, F. (2003) Vertical structure of the phytoplankton community associated with a coastal plume in the Gulf of Mexico. *Mar. Ecol. Prog. Ser.*, **251**, 87–101.

Weeks, A. R., Fasham, M. J. R., Aiken, J., Harbour, D. S., Read, J. F. and Bellan, I. (1993) The spatial and temporal development of the spring bloom during the JGOFS North-Atlantic Bloom Experiment, 1989. *J. Mar. Biol. Assoc. U.K.*, **73**, 253–282.