Title
Seasonal dynamics and interactions among Baltic Sea prokaryotic and eukaryotic plankton assemblages

Running title
High-frequency planktonic dynamics and interactions

Authors
Luisa W. Hugerth (1), Markus V. Lindh (2), Conny Sjöqvist (1), Carina Bunse (2), Catherine Legrand (2), Jarone Pinhassi (2), Anders F. Andersson (1)

(1) KTH Royal Institute of Technology, Science for Life Laboratory, School of Biotechnology, Division of Gene Technology, Stockholm, Sweden.
(2) Centre for Ecology and Evolution in Microbial model Systems - EEMiS, Linnaeus University, Barlastgatan 11, SE-39182 Kalmar, Sweden.
Abstract

One of the main goals of microbial ecology is to identify the mechanisms that regulate patterns in community structure at temporal scales compatible with populations' turnover times across complete seasonal cycles. Here, we examined high-frequency temporal dynamics of marine plankton from a sampling effort covering 2011-2013, roughly twice weekly, comprising 144 samples. Bacterial and eukaryotic communities were profiled by 16S and 18S high-throughput sequencing, respectively. Interestingly, we found that no combination of the measured environmental parameters could predict a significant proportion of the variation in population dynamics of bacterioplankton, and even less so for eukaryotic plankton. Large differences in physicochemical conditions and community composition typical of temperate climates mean that different regimes can quickly succeed each other over the year, with the relative importance of different drivers changing equally rapidly. Nevertheless, our approach revealed interesting recurrent co-occurrence patterns across distinct environmental changes. Hence, we could make abundance predictions for more than half of the most frequent OTUs based on interactions with other OTUs. These results suggests that a complex set of biotic interactions are contributing to temporal patterns among planktonic assemblages despite rapid changes in environmental conditions.

Introduction

A fundamental quest in microbial ecology is to identify which biotic and abiotic factors are regulating shifts in community structure. One important aspect of this is to determine whether microbial communities are mainly shaped by interactions between bacterial lineages, across domains (such as with viruses or protists) or by changes in environmental conditions. Ideally, this issue should be addressed by profiling natural communities over several years and with sampling intervals compatible with the time-scale of population turnover within a community, typically 3-5 days (Noble and Fuhrman, 2000).

Recent developments in high-throughput DNA sequencing allow for detailed temporal investigations of complex assemblages of microorganisms (Poisot et al., 2013). Seasonal variation in environmental conditions drives a plethora of bacterioplankton community responses to changes in environmental conditions (Fuhrman et al., 2006; Andersson et al., 2010; Gilbert et al., 2012; Lindh et al., 2015). Still, most studies have
measured community shifts and the impact of changes in environmental conditions at monthly intervals (Fuhrman et al., 2006; Andersson et al., 2010; Gilbert et al., 2012) or have been limited in their duration (Lindh et al., 2015). Nevertheless, high temporal resolution analysis revealed detailed bacterioplankton population dynamics in the Baltic Sea Proper indicative of a wide spectrum of niche differentiation patterns within each major bacterial group. Lindh and colleagues (Lindh et al., 2015) observed a relatively smooth continuous succession of bacterioplankton community shifts within specific seasons, but a pronounced and rapid shift between seasons. Similarly abrupt shifts in community composition and diversity have been observed in the San Pedro basin, the English Channel and off the Californian coast (Hatosy et al., 2013; Needham et al., 2013) where distinct shifts in bacterioplankton population dynamics and beta diversity occurred at both short-term and long-term temporal scales. Nevertheless, several gaps remain in our knowledge and understanding of bacterioplankton populations regarding the drivers of temporal dynamics, interactions with eukaryotic plankton and potential responses to environmental perturbations occurring at high-frequency over interannual time scales.

Despite the fundamental ecological role of eukaryotic phytoplankton communities at the base of aquatic food webs, the processes behind temporal changes in community structure are not well understood (Olli et al., 2013). Still, decades of observations have pointed out some of the factors that are likely involved. The temporal turnover in phytoplankton and protist biomass and species composition in the Baltic Sea is in general dictated by several interacting factors such as inorganic nutrients, temperature, stratification, light availability, zooplankton grazing and sinking rate (Harrison et al., 1986; Schiewer, 1988). A typical seasonal succession begins with the productively most important spring bloom, dominated by N-limited diatoms (Bacillariophyceae) and dinoflagellates (Dinophyceae) (Tamminen and Andersen, 2007). In the Baltic Proper, the spring bloom is usually initiated in the end of March-beginning of April (Wasmund et al., 1998). Immediately after the spring bloom, different species of hetero- and mixotrophic ciliates (Ciliophora) reach maximum abundances (Johansson et al., 2004). This period is also characterized by mesozooplankton, for example rotifers (Rotifera) such as Synchaeta spp. and Keratella spp., which may constitute up to 60% of the total zooplankton community in June (Johansson et al., 2004). When water temperatures reach +16°C in July-August there is another shift in planktonic composition. This increase in water temperature, together with calm weather, usually sets off massive P-limited cyanobacterial blooms in the Baltic Proper.
(Wasmund, 1997) that are able to fix nitrogen from the atmosphere, giving them a competitive advantage compared to N-limited autotrophs (Tamminen and Andersen, 2007). The decaying cyanobacterial blooms are a nitrogen source for picoplankton and eukaryotic protists (Ploug et al., 2010; Wannicke et al., 2013). Decreasing temperatures after the summer bloom, together with an increased frequency of storms, relaxes the water column stratification and stocks of inorganic nutrients in the surface waters are replenished. Thus, in September-October a peak in diatom abundance is usually observed (Wasmund and Uhlig, 2003).

Here we present the results of a three-year long (2011-2013) sampling effort from an offshore station in the Baltic Proper, comprising a total of 144 samples. Environmental parameters, microscopy data and 16S-based profiles of the bacterial community were collected for all three years, and two years also include 18S-based profiling of eukaryotic plankton (2012-2013). The frequency of sampling, of up to twice a week during the most productive period, is compatible with the population doubling times in this station (Lindh et al., 2015) and estimated community turnover times in marine environments (Noble and Fuhrman, 2000), which should prevent loss of signal between sampling events.

Materials and Methods

Sampling, DNA extraction and sequence library preparation

Unfiltered natural seawater was collected nearly twice weekly at the Linnaeus Microbial Observatory (N 56°55.851, E 17°03.640), at 2 m depth, depth using a Ruttner sampler, during the productive season 2011-2013. The absolute depth at the site is 43 m and it is located 10 km off the east coast of the Island Öland in the Baltic Sea Proper. 1 L of water was collected in the 2011 samplings and 10 L the following years. Water was stored in acid-washed, MilliQ-rinsed polycarbonate bottles and transported back to the laboratory within 1 h. Physicochemical parameters and microscopy data were collected as described in Lindh et al (Lindh et al., 2015). 16S amplicons were amplified with primer pair 341f-805r as described in Hugerth et al (Hugerth et al., 2014b) and 18S amplicons with primer pair 574*f-1132r as described in Hugerth et al (Hugerth et al., 2014a). All sequencing was done on a MiSeq instrument (Illumina Inc, Carlsbad, CA) at NGI/SciLifeLab, Stockholm.
Sequencing data processing and annotation

To eliminate low quality bases, all samples were trimmed to 230 bp forward reads and 170 bp reverse reads. Concatenated sequences with more than 2 expected errors were discarded. DADA2 (Callahan et al., 2015) was used to construct a sequencing error model, correct errors and eliminate chimeras from each sequencing run; a total of 5 runs were conducted for the 16S dataset and 2 runs for the 18S dataset. Errors were modeled from a Loess error function in self-consist mode based on up to 30 samples per sequencing run. Because the same correct sequences were inferred from multiple runs, the corrected reads from each DADA2 analysis for each gene were concatenated and dereplicated. All reads for each gene were mapped to this set at a minimum of 99.5% identity with USEARCH (Edgar, 2013) to build the final OTU table. At this point, 5,780-256,825 reads were left for each sample in the 16S dataset and 5,006-223,216 reads for the 18S dataset.

18S reads were annotated based on the PR2 database (based on GenBank 203) (Guillou et al., 2013) as described by (Hu et al.). 16S reads were annotated with the RDP database release 11 (Wang et al., 2007), SINA/Silva release 123 (Pruesse et al., 2012) and by blasting (v.2.2.28+) (Camacho et al., 2009) against the sequences presented by Newton et al (Newton et al., 2006) and the PhytoRef database (v.1.1) (Decelle et al., 2015), using the same strategy described by Hu et al. 3.5-99% of the eukaryotic reads were classified as metazoan (average 63%) and discarded from downstream analysis. Samples with less than 5000 remaining reads were discarded.

Statistical analyses and network calculation

Unless otherwise stated, statistical analyses were performed in R v.3.2.2 with package Vegan v.2.3-3 (Oksanen et al., 2008). Prior to variation partitioning, all variables were Hellinger-transformed. Metadata was assessed for redundancy online using Gustame (Buttigieg and Ramette, 2014) to remove factors with a spearman correlation >0.7. Networks were built based on Spearman correlations using package ccrepe v.1.7.0 (Schwager et al., 2014). Network visualisation and network property calculation were performed in gephi v.0.9.1 (Bastian et al., 2009). For most analyses, OTU present in at least 10% of the samples or ever reaching a relative abundance >0.5% of a single sample were kept and each sample was normalised to 1. However, to reduce noise from low frequency taxa, variation partitioning was done on OTU was present in >=25% of samples and network analyses on OTU present in >=50% of samples, and each
sample was Hellinger transformed. When comparing the seasonality of discrete OTU pairs, only pairs present in at least 10 samples of the relevant years were kept. The significance of correlation between metadata and sequencing data was assessed by anova on bootstrapped redundancy analysis for individual factors or a Mantel's test for matrices (2999 permutations per test). For the random forests procedure, up to twenty features were selected per target using minimum Redundancy Maximum Relevance filter on R library mRMRe v.2.0.5 (De Jay et al., 2014). The forests were then generated using package randomForest v.4.6-12 (Breiman et al., 2015). When comparing measured data with randomized data, the predicted data was shuffled 20 times and the average RMSD between the randomized data and the measured data was considered.

**Results**

**Community composition and seasonal patterns**

21,107 prokaryotic (16S) OTUs and 1932 eukaryotic (18S) OTUs were detected. Most of these are present only occasionally, with 57% of prokaryotic OTUs and 50% of eukaryotic OTUs present in less than 10% of samples. Only 13% of each set of OTUs are present in half the samples or more, but these correspond to over 87% of all reads in each dataset. Only 2 eukaryotic OTUs and 13 bacterial OTUs are present in every sample. Alpha diversity is generally constant, with a tendency to fall in late spring (fig. 1). Throughout the entire sampling period, the bacterial community was dominated by Actinobacteria (27-70%), Bacteroidetes (1.7-61%) and Cyanobacteria (1.7-74%). The non-metazoan eukaryotic community is by far dominated by Alveolata (18-94%), with high abundance of Ciliophores (1-32%) and Archaeoplastida (1-54%) as well (fig. 2).

Despite the low abundance of most OTUs, 44 bacterial tags and 57 eukaryotic tags occasionally reach over 5% of the total number of reads. All of them recur in all years and occasionally reach abundances >1%. In general, there is an exponential relation between the frequency of an OTU and its average abundance in the dataset ($R^2 > 0.78$) (fig. 3).

16S tags present a clear and recurrent seasonal pattern (fig. 4). 18S tags have a somewhat less clear presentation, but still present spring/early summer community somewhat separated from the fall community. These patterns can also be quantified at the individual OTU level: 60% of all OTUs present their maximum abundance in the same meteorological season each year, and less than 40% present their maximum abundance in the same relative time-frame to the spring and
summer algal blooms. The clear seasonal pattern is mostly driven by abundant OTUs: 73% of the 100 most abundant OTUs peak in the same season each year, and have dynamics largely independent of blooming events. There is an indication that a combination of seasonality and community drift is at play: 83% of OTUs present Spearman’s correlation >0.5 between each pair of consecutive years, and 73% percent present it for both. However, only 63% of OTUs have a Spearman correlation > 0.5 between 2011 and 2013, and only 49% over all three years.

The same general trend of low similarity across years is noticeable when considering correlations between prokaryotic OTUs. At p and q-value cutoffs of 0.05, 30-40% of links are conserved between each pair of years, and these values remain largely unaltered by decreasing p and q to as low as 0.001, after which the degree of concordance drops. Higher concordance between years is obtained when considering interactions predicted not on individual OTUs, but on higher order clustering of OTUs to the level of genus, family, order or class, although the total number of nodes and links decreases for each of these networks. At all levels from individual OTU to order, about 70% of the links predicted for each year were also identified when considering all samples to build the network. Eukaryotic tags form up to hundreds of times less links than prokaryotic, and only about 10% of links are common between 2012 and 2013.

Conditionally rare taxa

206 bacterial tags and 210 eukaryotic tags reach abundances >1% of the total sample on at least one occasion. Of these, 24 bacterial and 13 eukaryotic tags have coefficients of bimodality >0.9 and are therefore classified as conditionally rare (Shade et al., 2014). 11 of the 24 bacterial CRT present two peaks, one on April 20th 2011 and another on May 26th 2011. Many of these also present a smaller peak on May 27th 2013, but none in 2012. They are split between 4 bacterial classes (Flavobacteria, Alpha-, Beta- and Gammaproteobacteria) and are most likely heterotrophs, though some might be capable of mixotrophy. While the April 20th peak coincides with a peak in total phytoplankton biomass driven by diatoms and skeletonema cells, no measured parameters account for the May peaks of these taxa. In contrast, only 2 of the 13 eukaryotic CRT coincide in their single peak. Further, while only two of the bacterial CRT are autotrophs (one photosynthetic Anabaena and one chemoautotrophic Nitrospira), at least 7 of the 13 eukaryotic CRT are photosynthetic, while 3 are Alveolata whose trophic mode could not be assigned.

Despite representing only 0.1% of the prokaryotic taxa, bacterial CRT account for up to 66% of the Bray-Curtis community dissimilarity
between samples. Similarly but less dramatically, eukaryotic CRT make up 0.6% of the total number of OTU and account for up to 21% of the Bray-Curtis community dissimilarity.

Seasonality is the strongest predictor of bacterial community composition

Overall, 13% of the shifts observed for bacterioplankton community composition in the 16S dataset can be explained by changes in water temperature alone, which increases to 20% when including day length. The most important inorganic nutrients are phosphate and nitrate, which accounts for 7% and 3.5% of the variation, respectively (RDA, p<0.001 for both), but is highly redundant with the other two parameters. A combination of water temperature, day length and DOC concentration explains 22% of the variance (p<0.001). Less than 1% of the variation in the bacterial dataset can be explained by total phytoplankton counts or phytoplankton biomass, although both parameters present highly significant correlations with the data (p<0.001).

The eukaryotic dataset has more unexplained variation: 10% of it can be attributed to temperature alone and 13% to a combination of temperature and day length. While phosphate and bacterial production show highly significant correlations (<0.001) with the data, they don't add further explanatory power to the model. The exception to this is eukaryotic phytoplankton, 12% of the variance is ascribed to temperature, and 16% can be explained by a combination of temperature, day length and phosphate (p < 0.001). Heterotrophs and mixotrophs have other highly significant correlations, most markedly with dissolved organic carbon and bacterial production, respectively; however, neither of these variables adds to the predictive power of temperature and day length alone.

Comparing distance between samples based on different measurements (Mantel's test over samples), neither distances between eukaryotic profiles nor between bacterial profiles display significant correlations with the physicochemical parameters measured (p>0.5). However, they do present weak but significant correlations with each other (p<0.001, Pearson's r=0.35) and with microscopy data (16S: p<0.001, r=0.16; 18S: p=0.002, r=0.14).
Drivers of community composition change over the year

While temperature and day length are main drivers of bacterial community change, they’re not the only ones. However, in analysing over the course of several years, important seasonal variability in environmental conditions may have been missed. Therefore, the time-series for each year was split into five relevant periods: pre-bloom, spring bloom of diatoms and dinoflagellates, inter-bloom, summer bloom of filamentous cyanobacteria and post-bloom. The bloom periods for each year were defined on the basis of microscopy data. The time-points belonging to each of these periods in all years were concatenated and subjected to variation partitioning. In the pre-bloom, silicate is the single strongest explanatory variable ($r = 0.088$, $p = 0.012$). This variable, however, is highly redundant with other environmental factors during this period, possibly suggesting that changes associated with the spring bloom are in course before the bloom itself becomes visible. The best explanatory model for this period, albeit still weak, is a combination of water temperature, nitrate and phosphate ($r = 0.228$, $p = 0.013$). During the spring bloom, silicate in itself is not a strong community driver ($p = 0.12$). However, it does contribute with increased explanatory power to significant variables (temperature, day length, phosphate, total nitrogen and dissolved organic carbon). The best overall models for this period consider temperature, day length, dissolved organic carbon and either phosphate or silicate ($r = 0.209$ and $r = 0.223$, respectively; $p = 0.0003$ for both). During the short period between blooms, temperature and day length alone can explain upwards of 35% of the variation ($p=0.001$ for both), and combining them covers 63% ($p=0.0003$). Finally, considering total phytoplankton counts in addition to light and temperature accounts for 75% of the variation ($p=0.0003$). This high predictive power is already lost during the summer bloom, when water temperature and day length can only account for 21% of the variation ($p = 0.0003$). However, the predictive power is restored when taking into account the degree of bloom, by adding either the role of chlorophyll a alone ($r = 0.299$, $p = 0.001$) or chlorophyll a and dissolved organic carbon ($r = 0.330$, $p=0.001$). After the summer bloom is over, while many factors present significant correlations to the community, these are generally very weak (temperature, day length, silicate, ammonium, nitrate, total nitrogen and phosphate). The best model, a combination of temperature, day length, ammonium and silicate, is highly significant but only weakly explanatory ($r = 0.201$, $p=0.0003$).

The eukaryotic community has in general a smaller proportion of variance explained, but enough to suggest that different mechanisms are at play. Before the spring blooms, nitrate and DOC are the only statistically
significant drivers of the community \((r=0.107, p=0.015; r=0.37, p=0.024)\). Together they can account for 16% of the community variance. However, the best explanatory model is a simple combination of temperature and phosphate \((r = 0.229, p=0.012)\). During the spring bloom, temperature and DOC are significantly correlated to the overall community \((r = 0.086, p = 0.007; r = 0.111, p = 0.0007\), respectively). A combination of the two accounts for 14% of the variation \((p = 0.001)\), which is increased to 17% when accounting for day length \((p = 0.0003)\). During the summer bloom, predictability is still low, with only total nitrogen presenting significant correlations to the community \((p = 0.045)\). However, the best predictor for this period is a combination of temperature, day length and silicate, which is somewhat improved by the addition of nitrate \((r = 0.167, p = 0.006; r = 0.182, p = 0.035, \text{respectively})\). Finally, in the post-bloom period, predictability is completely lost. The only significantly correlated parameter is total nitrogen \((r = 0.08, p = 0.04)\), and no combination that accounts for more than 10% of the variance attains statistical significance.

Predictive models for individual OTU

Another approach for finding predictors for individual OTU is fitting models through random forest regression. For 752 (506 bacteria, 246 protists) OTUs present in at least half the samples for which there is both 16S and 18S data, we trained random forests to predict the abundance of the OTU based on either abundances of other OTUs or environmental data, or both. Since the 2012 data was found to be less predictive (see below and (Legrand et al., 2015)), models were trained on 2013 data only. 429 (309 bacterial, 120 protist) OTU could be fitted with a model with \(R^2 > 0.5\). Of the bacteria, 31% had their best model formed by combinations of other bacterial taxa only, while 59% were better resolved by a combination of bacterial and protist taxa, 2% by a combination of bacterial taxa and environmental parameters and 8% by all three data sources. In the case of protists, 17% are best predicted by other protists, 72% by a combination of bacteria and protists and 11% by a combination of all data sources.

In addition to this assessment by bootstrapping, it is also possible to assess how well models built in one year fit data from the following. There are no fewer OTU with a model fitted from one year and tested in either of the others compared with testing it on the same year. Root-mean-square deviation (RMSD) between predicted and measured data is significantly different from that from randomized data for all models built based on 2011 data and for models built in 2013 and tested in 2013 and 2011 (paired Student's t-test, \(p<0.0001\)). The 2012 model is not different
from random even when tested on 2012 data. This is in agreement with the weaker seasonal signature observed for that year (fig. 1).

Finally, due to fitting our models on a time-series, we could also assess whether a model could be used to predict the relative abundance of an OTU in the future. This was done only for the bacterial data, since the eukaryotic data had missing samples which made the sampling interval irregular (4-20 days). In this case, we observed that while much fewer bacterial OTU could be predicted ahead of time, these were still highly significant. For 83 bacterial tags, models could be found to predict their behaviour in one time point based on the previous one, 61 could be predicted two time points ahead and several others (20-40 per interval) could be predicted at longer time-spans (fig. 5).

**Discussion**

Here we present a highly resolved, three-year long time-series of surface waters in the Baltic Sea Proper. Despite the sequencing depth (on average 52,000 reads), only 13 bacterial OTUs and 2 eukaryotic OTUs were found in every sample, meaning that even persistently abundant taxa can occasionally fall below detection limit. This could potentially be explained by occasional and unpredictable blooms of rare taxa taking up a large portion of the sequencing capacity, but no correlation between blooms of conditionally rare taxa and Shannon diversity was observed. Despite the low detection frequency of most OTUs, we observed repeatability of the overall community structure and dynamics by a number of measures, including the season of maximum frequency for each OTU, its relation to blooming events and its correlation with other OTUs. For some of these measures repeatability appears equally large across all pairs of years, while for others it seems larger between non-consecutive years. This is probably due to unusual environmental conditions with a mild winter in 2011-2012 (Legrand et al., 2015). A longer time-series, perhaps spanning up to a decade, would be needed to assess whether there is consistently lowering similarity between a year and following ones, and at which point this trend saturates.

Few strong correlations were found between bacterial and eukaryotic taxa. This phenomenon can have two non-exclusive explanations, one technical and one biological. On the technical side, 16S and 18S tags were not amplified in the same reaction, so the relative proportion of them cannot be assessed. While care was taken to keep in vitro and in silico conditions as close as possible, and Spearman’s rank correlation was chosen because it should be less sensitive to this, it is still
possible that signal was lost in the process. Another unavoidable issue is that aquatic protists can have highly variable and wildly different copy numbers of ribosomal RNA genes (Gong et al., 2013), so that if two clones with nearly identical 18S rRNA tags and biological function succeed each other more or less stochastically, the result will be dramatic shifts in the apparent abundance of this eukaryotic tag, effectively decoupling it from associated 16S tags. However, it is also possible that a large proportion of prokaryotic-eukaryotic interactions is not specific at the OTU-level. Grazing, for instance, is often deemed to be based mostly on the size of the prey (Montagnes et al., 2008). Other studies have found that, while highly specific and ecologically relevant links exist between eukaryotic and prokaryotic taxa, the majority of significant correlations are found within domains (Steele et al., 2011; Lima-Mendez et al., 2015).

Various unmeasured factors could potentially confound the drivers of community structure in this study. For instance, life cycle dynamics in dinoflagellates, diatoms and ciliates, which have been observed to include biennial cycles, could impact the abundance of specific taxa from year to year (Wyatt and Jenkinson, 1997; D’Alelio et al., 2010). This may also in part explain the generally smaller proportion of variation explained compared to prokaryotes in our analyses. A growing number of taxa have been shown to have both benthic and planktonic life stages (Montresor et al., 2006) and quantifying only one of these may complicate the interpretation of their abundance in the pelagic zone. The predictability of eukaryotic phytoplankton occurrence and abundance is complicated by the fact that the transition between planktonic and benthic life stages of a species is controlled by endogenous factors, in conjunction with the physical and chemical environment (Anderson, 1998).

The best explanatory model for the spring bloom community included temperature and phosphate, which is somewhat surprising since the most dominating community members are regarded as nitrogen limited and silica being the most essential factor for diatoms (Tamminen and Andersen, 2007). However, the relatively small proportion of strict autotrophs and a high prevalence of mixotrophs in the studied community may explain the seemingly low importance of nitrogen and silica in this analysis. The higher abundance of potentially mixotrophic dinoflagellates compared to diatoms during the spring bloom is in concordance with recent findings of a shift in community composition across the Baltic Sea. The shift towards more dinoflagellate-dominated spring bloom communities during the last 30 years is linked to climate-driven mechanisms indicating that mild winters and thin ice cover is unfavorable for diatoms (Klais et al., 2013). The significant correlation of DOC to community structure during the spring
bloom also reflects the large proportion of non-autotrophic components. The predictability of environmental drivers during the summer bloom was low, which most likely results from the large proportion of mixo- and heterotrophic groups which are more driven by biotic factors and prey availability. The loss of predictability after the summer bloom is likely a result of mixing events resulting both in more stable environmental conditions and an increased displacement of community members in relation to their optimal habitat.

When considering only interactions between bacterial OTUs, 30-40% of links are shared between years. This is true at p-value and q-value cutoffs varying from 0.05 to 0.001, despite a 6-fold reduction in total number of links. Comparing each year’s network to one generated from all data, about 70% of links in each yearly network are also found in the full network, which contains 2-3 more links than the yearly networks. It is likely that, by including more time points, the full network has increased sensitivity, since simulation studies show that sensitivity of detection of Lotka-Volterra-like interactions by co-occurrence models increases up to 100 samples without a concomitant decrease in specificity (Berry and Widder, 2014). However, it is also possible that some of the interactions captured by the full network are real but transient, only realising their potential under appropriate conditions which were not met in every year studied. Interestingly, generating networks for clusters of higher taxonomic order greatly decreases the number of edges (from thousands to hundreds or less), but also increased the concordance between years to 40-60%, possibly indicating that different, more or less closely related, OTU can perform the same ecological functions in different years. However, the concordance between each year’s network and the full network remained stable at 70%.

The measured environmental parameters perform poorly in describing the OTU dynamics. Since thousands of OTU have been detected, each with its unique niche, perhaps only a detailed environmental survey including hundreds of micronutrients and macromolecules could be sufficiently predictive. It is not unusual that a study focused on a single site or relatively similar sites cannot resolve drivers of community change with higher resolution (Hatosy et al., 2013; Cram et al., 2015; Lima-Mendez et al., 2015), suggesting that high r-values found partitioning variation from highly unequal sites may be masking several unmeasured confounding factors. One set of taxa which highlight the degree to which drivers of microbial growth and demise are still unknown are conditionally rare taxa, which grow to make up a large fraction of the community before falling below detection levels again. Since the majority of bacterial CRT detected
are heterotrophs, it is possible that they thrive attached to particles, and samples on a few occasions contained more of their preferred kind of particle. This is congruent with the fact that 11 of 24 bacterial CRT have their maxima at the same time point. However, this is, to the best of our knowledge, the first exploration of conditionally rare protist taxa, and of the 13 identified, at least 7 are photosynthetic, including a parasitic member of order Syndiniales.

In addition to the life cycle of eukaryotes discussed above, other confounding factors may include allelopathic interactions, which have been shown to result in unexpected community structure (Fistarol et al., 2004). While these are partially taken into account by tracking eukaryotic communities, which will have important effects in the composition of dissolved and particulate organic matter, no direct measures were made of these. Another important driver of plankton turnover are viral communities, which were not tracked in this study but have been shown to affect both community and within-population dynamics in planktonic environments (Rodriguez-Brito et al., 2010; Parsons et al., 2012). Also, in compliance with the monopolization hypothesis, taxa that arrive first or taxa that are able to establish early on during the succession, will suppress later taxa through priority effects (Sefbom et al., 2015). This may result in randomized year-to-year abundances of different OTUs.

Despite all of these complicating factors, we have demonstrated with the bacterial dataset as a model that predictive models for at least frequent and abundant OTU are possible. This opens up the possibility of rationalizing environmental monitoring efforts by doing an initial high-frequency sampling for model building, followed by a more sparse monitoring that uses models to predict data where it is missing.
Acknowledgements

We thank Anders Månsson and Kristofer Bergström for their knowledgeable and persistent sampling effort, and Sabina Arnautovic and Emmelie Nilsson for their careful processing of samples. It is also funded by the Swedish Research Council VR (grant 2011-5689) through a grant to A.F.A, as well as by Formas project ECOCHANGE (Strategic Grant for Marine Research) through a grant to C.L. and J.P. All computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). Sequencing was conducted at the Swedish National Genomics Infrastructure (NGI) at SciLifeLab in Stockholm.

Figure legends

**Figure 1:** Alpha-diversity of the (a) prokaryotic and (b) eukaryotic communities over the sampling period. Sampling dates in this and all figures are presented in format yymmdd. Dates are also coloured by month.

**Figure 2:** Taxonomic composition of (a) prokaryotic and (b) eukaryotic communities over the sampling period. (c) Eukaryotic community defined by putative trophic mode.

**Figure 3:** The frequency of each OTU is plotted against its total abundance. The size of each circle represents the maximum abundance of this OTU across samples. Colours represent taxa and have the same meaning as in figure 2 (a) and (b).

**Figure 4:** NMDS of 16S OTU samples reveals clear seasonal patterns, despite the presence of more noise in the 2012 samples. Triangles: 2011; Circles: 2012; Squares: 2013. Grey: January; dark red: March; bright red: April; orange: May; yellow: June; bright green: July; dark green: August; dark blue: September; light blue: October; purple: November; lilac: December.

**Figure 5:** The number of OTU with significant models built from random forests is plotted in dark red against the time-delay (in samples; interval between samples 4-10 days). Overlayed in green is the log10 of the probability that the models calculated perform better than random (paired Student's t-test).
References


Figure 1: Alpha-diversity of the (a) prokaryotic and (b) eukaryotic communities over the sampling period. Sampling dates in this and all figures are presented in format yymmdd. Dates are also coloured by month.
Figure 2: Taxonomic composition of (a) prokaryotic and (b) eukaryotic communities over the sampling period. (c) Eukaryotic community defined by putative trophic mode.
Figure 3: The frequency of each OTU is plotted against its total abundance. The size of each circle represents the maximum abundance of this OTU across samples. Colours represent taxa and have the same meaning as in figure 2 (a) and (b).
Figure 4: NMDS of 16S OTU samples reveals clear seasonal patterns, despite the presence of more noise in the 2012 samples. Triangles: 2011; Circles: 2012; Squares: 2013. Grey: January; dark red: March; bright red: April; orange: May; yellow: June; bright green: July; dark green: August; dark blue: September; light blue: october; purple: november; lilac: december.
Figure 5: The number of OTU with significant models built from random forests is plotted in dark red against the time-delay (in samples; interval between samples 4-10 days). Overlayed in green is the log10 of the probability that the models calculated perform better than random (paired Student’s t-test).