Project:	BONUS Blueprint
Deliverable:	4.3
Work package:	4

Deadline: Month 49, **Nature**: Report, **Dissemination level**: Public

Responsible contact person:
Anders Andersson, KTH Royal Institute of Technology, Sweden,
anders.andersson@scilifelab.se

Report on signature genes for environmental status diagnosis

Introduction

The BONUS Blueprint project aims at developing a framework for predicting the environmental status of a water body based on the composition of microbial taxa and encoded gene functions in its plankton community. The basis for this approach is twofold. First, microorganisms are the main catalysts for most biogeochemical processes in seawater, thus microbial community composition is expected to be correlated with environmental conditions. Second, microorganisms have different niches and are selected for by the prevailing environmental conditions. Thus, microbial consortia are shaped by their environment, but the environment is also shaped by them. In the BONUS Blueprint project we measure microbial community composition by either taxonomic or functional composition. Taxonomic composition will vary because different taxa have different niches and carry out different processes. Functional composition will vary because organisms adapted to different niches and carrying out different processes will encode different sets of functional genes.

The goal of **D4.3** is to examine the coupling between microbial community composition and environmental conditions, and to establish how we can use the microbiome data to predict environmental data. We show that the microbiome is correlated with environmental conditions, both at the overall community level, and at the level of specific genes and taxa. We further show that we can predict specific environmental conditions from the metagenome data. This is demonstrated here for parameters that can quite easily be measured *in situ* or by routine lab procedures (such as temperature, nutrient levels, etc.), but we argue that the same approach should also be applicable for predicting conditions or states that are more difficult to measure by standard methods. Finally, we discuss this proof-of-principle in a marine monitoring context.

Data and Methods

The analysis is based on metagenome data from 97 samples collected during two sampling cruises and by time-series sampling at a single-station (Figure 1). The “Transect” cruise dataset consists of pelagic samples collected in the summer of 2014 from 3 depths at 10 stations covering the complete salinity gradient (Skagerrak to Bothnian Bay). The “Coastal” cruise data consists of mainly coastal surface-water samples collected in the summer of 2015 and covering the mesohaline (Baltic Proper and Gulfs of Finland and Riga). The “LMO” sample set consists of surface-water samples from the Linnaeus Microbial Observatory (LMO) located 10 km east of Öland, sampled weekly or bi-weekly during the ice-free season of 2012 (Hugerth et al. 2015).

Microbial cells from each water sample were captured on a filter, and total DNA was extracted from the mixture of cells on the filture. Shotgun sequening libraries were prepared and Illumina sequncing was conducting, generating on average 42 million metagenome sequencing read-pairs (2 x 100 bp) per sample.

In order to quantify functional genes and taxonomic groups in the samples in a reproducible and relatively quick manner, the data was bioinformatically mapped onto a reference metagenome (BARM; Baltic Sea Reference Metagenome) described in Deliverable 4.2 of the BONUS Blueprint project**.** BARM consists of 6.8 million genes assembled from 81 metagenome samples (most of which overlap with the samples analysed here) that have been taxonomically and functionally annotated (Alneberg et al. 2018). BARM can be accessed via BalticMicrobeDB with a graphical user interface (https://barm.scilifelab.se/). For each sample we quantified taxa at different taxonomic levels (Domain, Class, Order, Family, Genus) and gene functions using widely adopted annotation schemes (COGs, PFAMs, Enzyme Commission numbers, etc.). We also predicted abundances of different metabolic pathways in the samples (KEGG modules) based on the composition of enzyme classes. We refer to the quantifications as taxonomic or functional profiles of the samples.

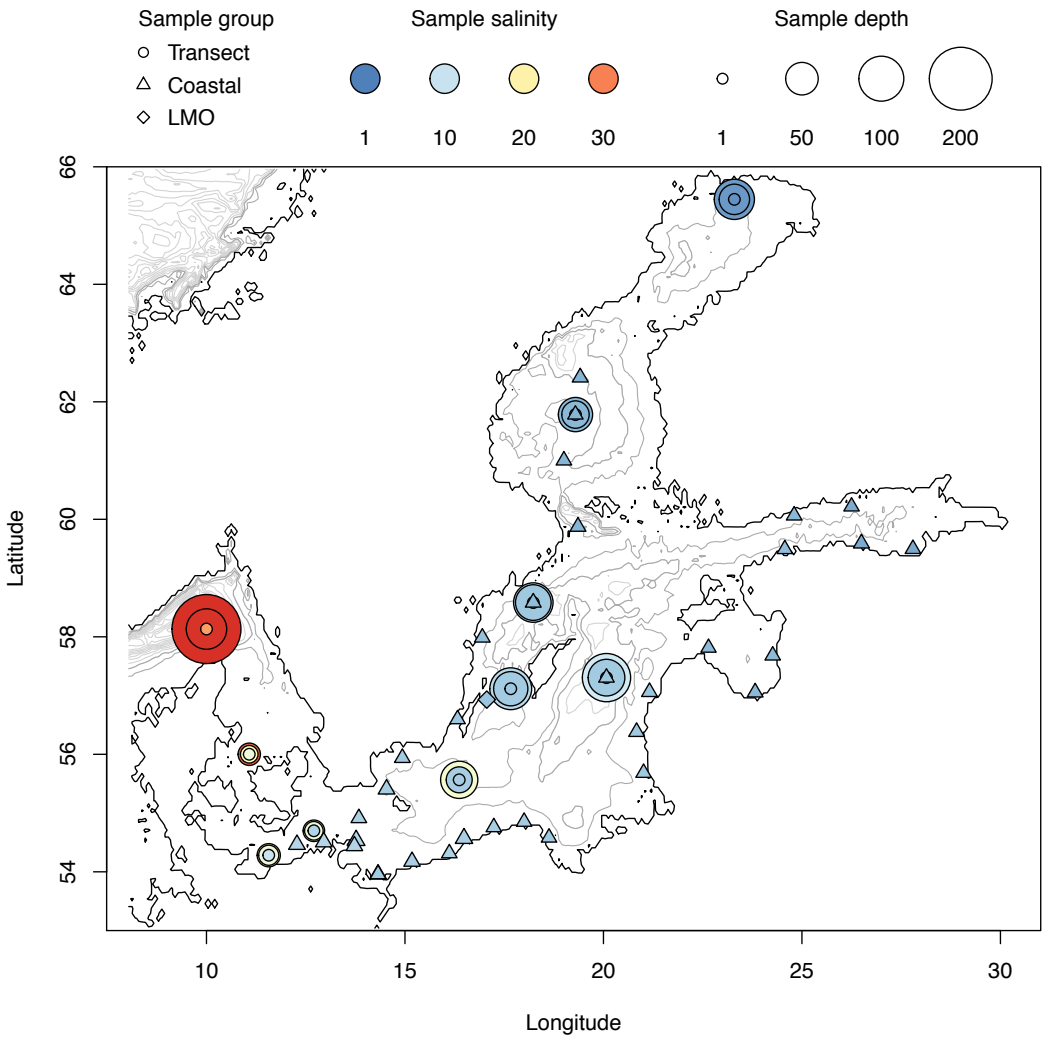

Figure 1. Map showing the sampling locations. The three sample sets included in the analysis presented here (Transect, Coastal and LMO) are indicated with different symbols. The marker colour indicates the salinity of the water sample while the size indicates the sampling depth. The contour lines indicate depth with 50 m intervals.

Results

Spatio-temporal variation of the Baltic Sea microbiome

Analysis of the samples from the Transect cruise reveals a gradual change in taxonomic composition along the north-to-south salinity gradient of the Baltic Sea, and also changes with depth (Figure 2a). The same dominant prokaryotic taxonomic groups were observed as in previous pan-Baltic studies (Herlemann et al. 2011; Dupont et al. 2014; Hu et al. 2016; Herlemann et al. 2016) and similar higher-level trends of an increase in *Alpha-* and *Gammaproteobacteria* and a decrease in *Actinobacteria* and *Betaproteobacteria* with increasing salinity. Thus, the overall structure of the Baltic Sea microbiome appears to be stable across

years. Seasonal dynamics is another characteristic of the Baltic Sea, which can also be observed in the metagenome data. As previously seen at station LMO (Lindh et al. 2015; Hugerth et al. 2015), succession in the phytoplankton community is paralleled by succession in the heterotrophic bacterioplankton community, with the major heterotrophic groups *Bacteroidetes* and *Actinobacteria* peaking in spring and summer, respectively (Figure 2b).

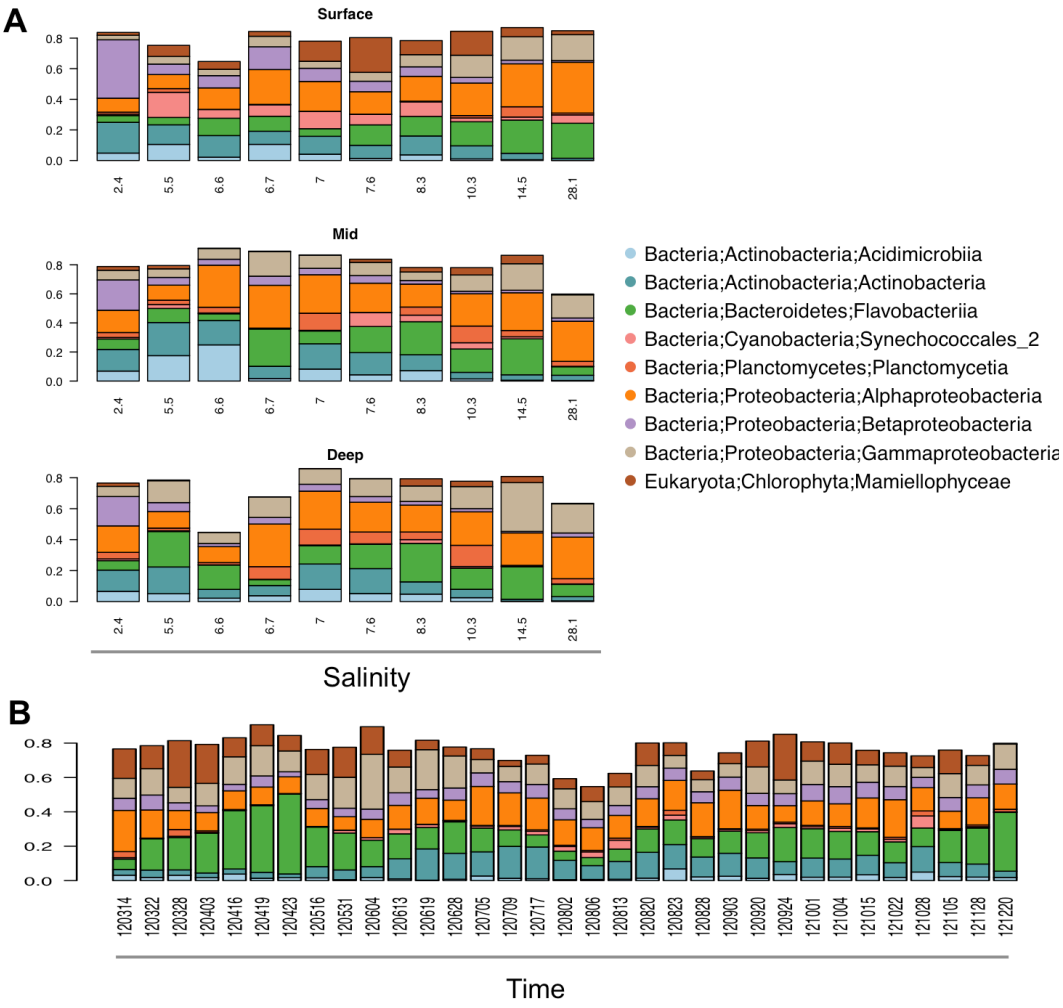

Figure 2. Barplots showing the relative abundance of dominant taxonomic classes (each representing > 0.01% of annotated reads) in the Transect and LMO datasets. **A**) Transect dataset with surface, mid and deep layer samples in separate barplots, going from low salinity (Bay of Bothnia) to marine (Skagerrak) conditions. Sample water salinity indicated below each bar. **B**) LMO (March to December 2012) samples, with sampling date indicated below each bar.

When instead analysing community composition based on functional data, we likewise see that samples are structured along the physio-chemical gradients of the Baltic Sea. Plotting the cruise samples in two dimensions based on their functional profiles (in this case counts of COGs, but other functional annotations give similar results) shows that functional community composition is correlated with depth, temperature and salinity (Figure 3a,b). Moreover, ordination of the LMO samples based on their functional profiles demonstrates a clear seasonal succession in the community (Figure 3c).

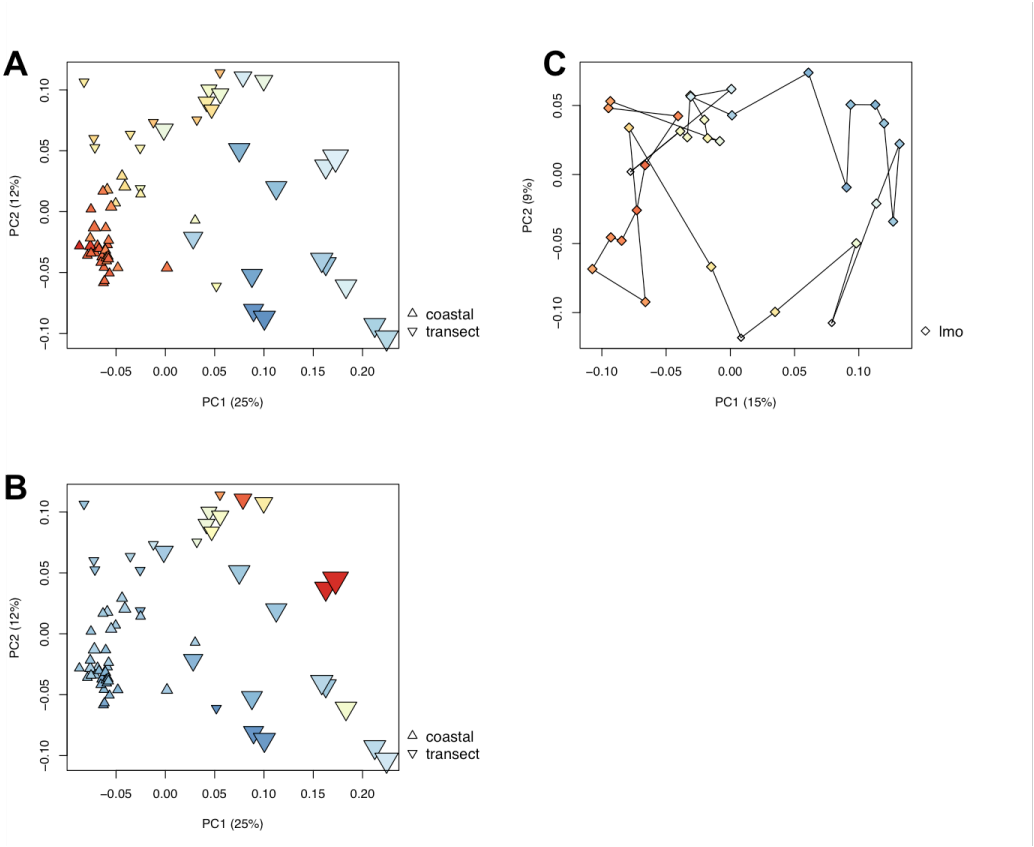

Figure 3. Principal coordinates analysis (PCoA) plots of community composition calculated as counts of functional genes (COGs). Samples with more similar community composition are more proximal in the plots. The percentage of variation explained by the principal coordinates (axes) are indicated. **A-B**) Transect data, with color representing temperature (in A) and salinity (in B), respectively, and size representing depth. **C)** LMO data, with color representing temperature, and samples adjacent in time being connected with a line.

Distribution of genes and pathways of biogeochemical relevance

The metagenome data also gives opportunities to see how genes for particular biogeochemical pathways are distributed in time and space. Microbes are main drivers of carbon, nitrogen and sulfur cycles in the sea. In Figure 4 we show how genes for three pathways (KEGG modules) for nitrogen cycling are distributed in the Transect dataset. Nitrogen fixation, the process of converting N_2-gas to ammonium, providing bioavailable nitrogen for bacterial and phytoplankton growth, is as expected mainly present in the surface layer since cyanobacteria are the primary contributors to nitrogen fixation in the Baltic Sea. It is also higher in Baltic Proper compared to the northern sub-basins, in accordance with the Baltic Proper plankton community being limited by nitrogen during summer. In contrast, nitrification (oxidation of ammonia to nitrate) and denitrification (oxidation of organic carbon using nitrate resulting in N_2 formation) are both restricted to deeper waters with lower oxygen levels. Interestingly, here we also detect genes for the dissimilatory nitrate reduction to ammonium (DNRA) pathway, although at ~20 x lower abundance than denitrification. DNRA is an alternative process to denitrification, where ammonia is produced rather than N_2, thus maintaining a bioavailalbe N source in the system. It

is occasionally a dominating nitrate sink in oceanic oxygen minimum zones but has been little studied in the Baltic Sea (Bonaglia et al. 2016).

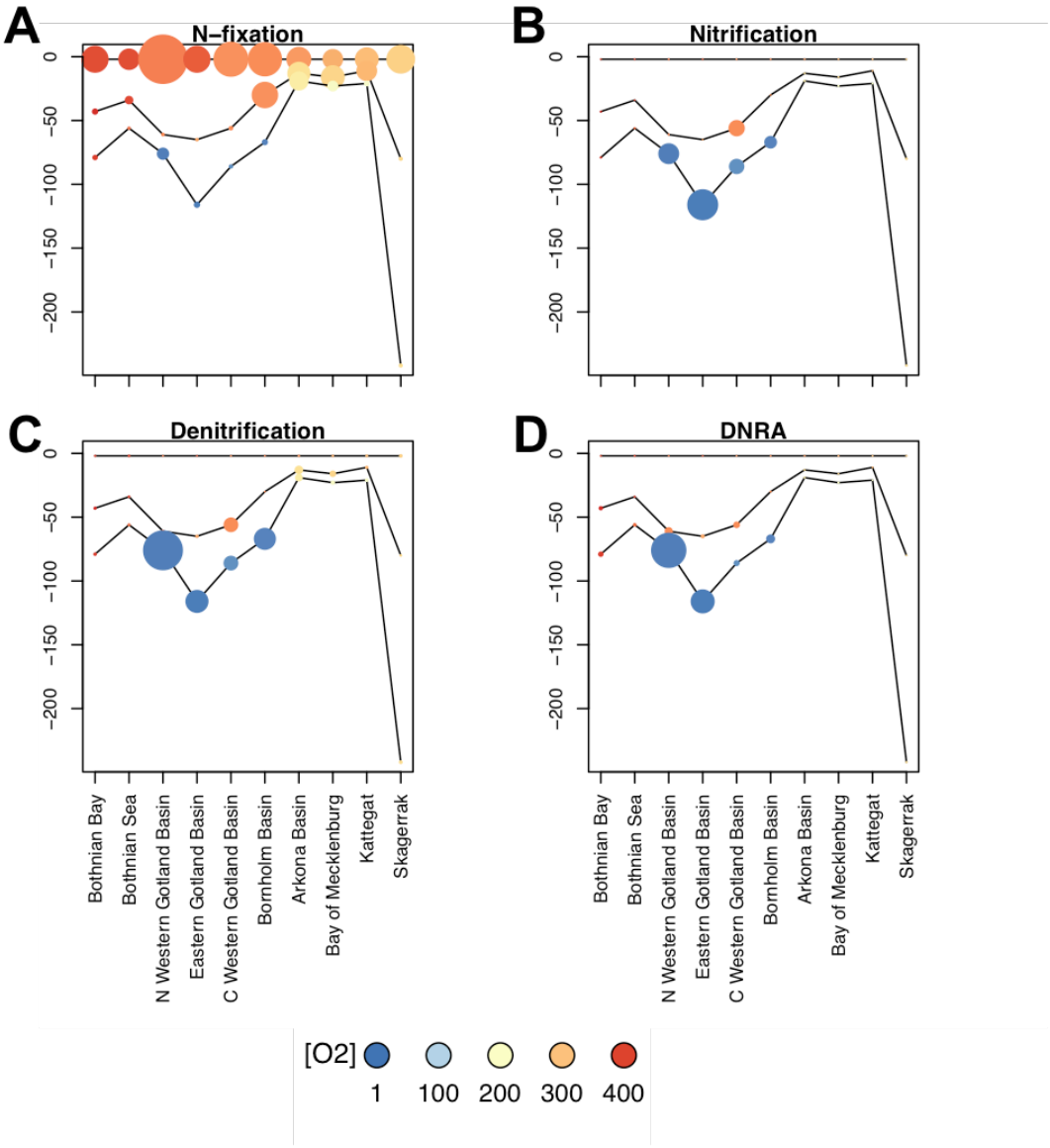

Figure 4. Relative abundance of genes for select nitrogen metabolic pathways in the Transect metagenomes. **A**) Nitrogen fixation (KEGG module M00175). **B**) Nitrification (M00528). **C**) Denitrification (M00529). **D**) Dissimilatory nitrate reduction to ammonium (DNRA) (M00530). Lines connect samples from the same depth layer (surface, mid or deep). Y-axis represents sampling depth and dot color oxygen concentration (µmol/l). Dot size represents mean abundance of the genes in the pathway. Different magnifications of the dots were used for different pathways; with 500, 750, 750 and 2500 x magnification in A, B, C and D, respectively.

Besides nitrogen, phosphorous (P) is an essential nutrient for sustaining plankton growth, being a key component in biomolecules such as DNA, RNA and phospholipids. Due to inflow of riverine water depleted of P, the plankton communities of the northern basins of the Baltic tend to be more P-limited than the southern basins. But there is also a temporal dimension with

excess amounts of nutrients in winter that is depleted during the spring phytoplankton bloom. The summer community of the central Baltic Sea is primarily N limited but also by P, especially during blooms of nitrogen fixing cyanobacteria. Microbes can take up inorganic phosphorous as phosphate ions either using the constitutively expressed low-affinity uptake system PitA/PitB or using the phosphate-starvation induced Pst system. In our metagenomes we see a gradual change in the abundance ratio between high- and low affinity system genes along the Baltic transect (Figure 5), with the highest ratio in the north, consistent with the microbes in these waters being adapted to more P-limited conditions than in the south, since high-affinity systems are better for taking up low-concentration ions.

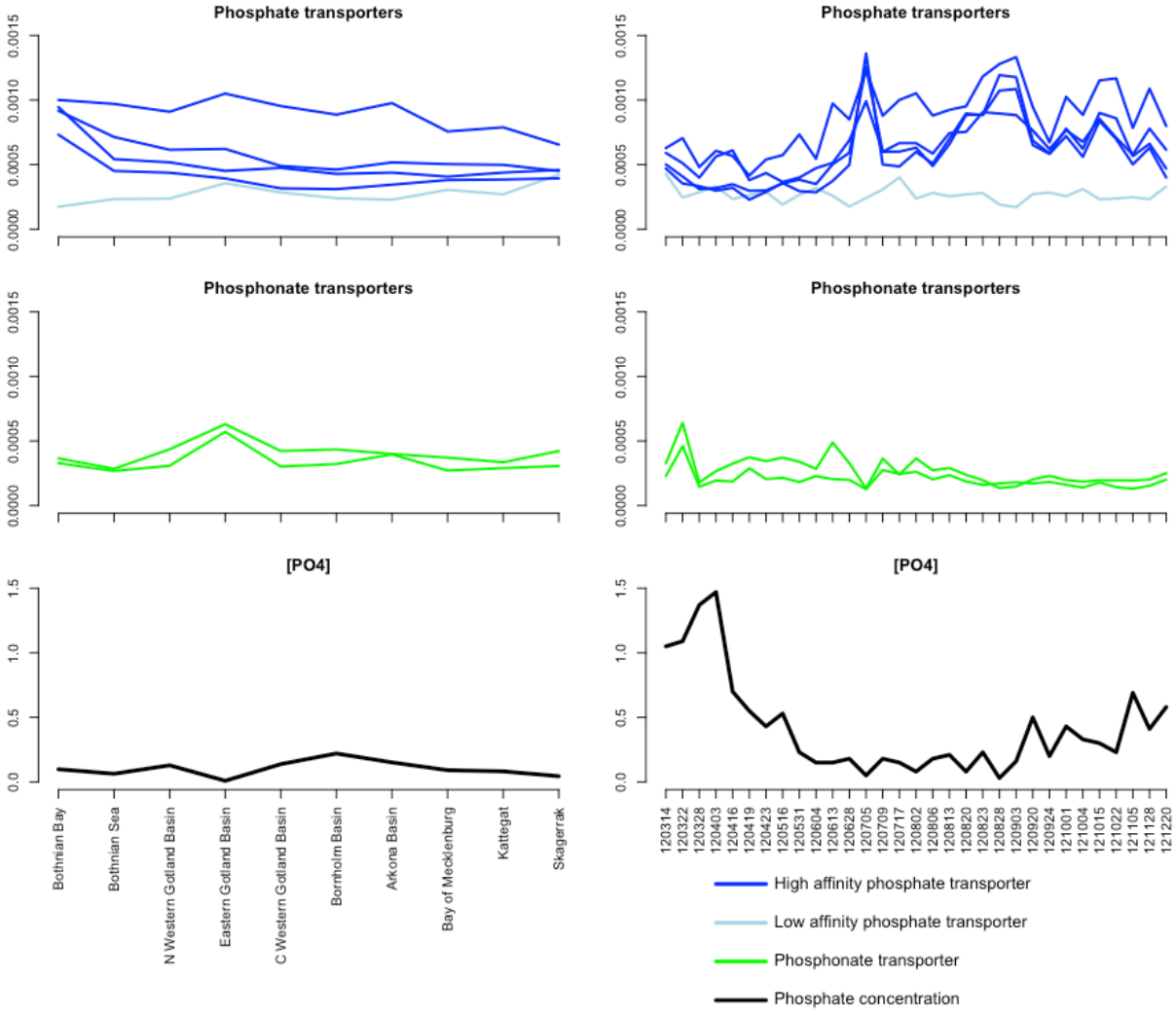

Figure 5. Relative abundance of gene families for phosphate and phosphonate uptake along the Transect (left panel) and at station LMO (right panel). The high-affinity phosphate transporter gene families (COG0581, COG1117, COG0573, COG0226) belong to the ABC-type transporter complex pstSACB and the low-affinity family (COG0306) to the PitA/PitB uptake system.

We also see a temporal trend at station LMO with a higher ratio after the spring bloom when P concentrations are lower (Figure 5). Microbes can also utilize organic phosphorous such as phosphonates, as recently demonstrated for *Nodularia spumigena* (Teikari et al. 2018). The BARM metagenome includes a large number of phosphonate uptake genes belonging to two major gene families (Figure 5). The abundances of these are usually lower but in the same order of magnitude as the inorganic phosphate uptake gene families, indicating that phosphonate may be an important phosphorous source for Baltic Sea microbes.

Detection of toxin genes

In addition to monitoring genes and metabolic pathways of biogeochemical relevance, the metagenome can be used to monitor features of relevance to human and animal health, such as toxin genes. The Cyanobacterium *Nodularia spumigena* is known to produce the hepatotoxin nodularin by utilizing the nodularin synthetase enzyme complex (Sivonen et al. 1989). The *nda* genes coding for the required enzymes are located in a major cluster in the genome. Koskenniemi et al. (2007) showed that the prevalence of the *ndaF* gene as assessed by quantitative PCR (qPCR) in Baltic Sea waters correlated strongly with concentration of the actual toxic compound nodularin. They concluded that *Nodularia* cells produce the toxin at a constant level during the period and locations studied. Therefore, the detection of *Nodularia* sequences harbouring *nda* genes in metagenomic data is likely to indicate relevant levels of the nodularin toxin in a given sample. We identified all *Nodularia nda* genes in the BARM metagenome (with average 99% amino acid identity to those of *N. spumigena* CCY9414). In Figure 6 we quantified these genes in the LMO data set. The presence of *nda* genes coincided with the presence of *N. spumigena* filaments in the water, as detected by microscopy. The same approach can potentially be used to monitor other plankton toxin genes.

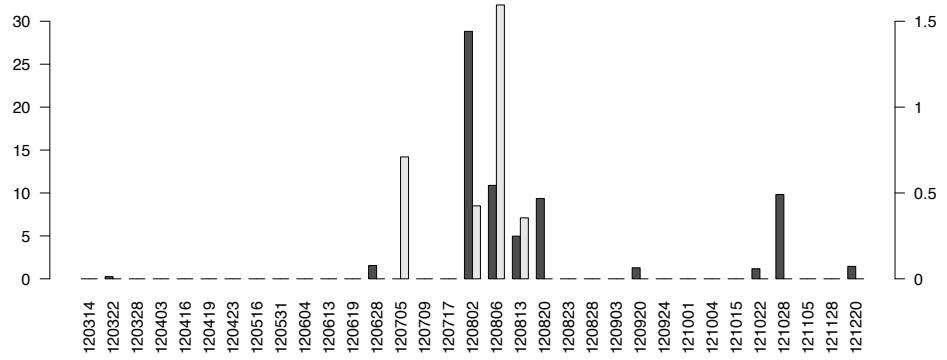

Figure 6. Barplot of *nda* gene abundances (sum of all *nda* [*ndaA* – *ndaI*] genes) (darkgrey bars, right y-axis) in the LMO metagenomes, and microscopy-estimated *N. spumigena* filaments per ml (lightgrey bars, left y-axis) for the same set of samples.

Predicting environmental parameters from the microbiome

Above we demonstrated how overall community composition and specific genes and pathways change with environmental conditions. Below we show that we can also predict environmental

conditions from the metagenome data. For doing so we use a machine-learning approach called Random Forest (RF) regression (Breiman 2001). In RF an ensemble of decision trees are constructed, each tree predicts the value of a response variable based on the values of a set of features. In our case the response variable is an environmental parameter such as temperature or phosphate concentration, and the features are relative counts of microbial taxa or functional gene groups. Each tree is trained on a different random subset of the samples, and using a random subset of the features. This procedure makes the predictions robust and also makes it possible to assess the prediction accuracy by evaluating the trees using the samples that were not included in the training of those trees (out-of-bag evaluation). RF also identifies the most important features for predicting a parameter (i.e. what taxa or functional genes).

We used the metagenome data from the 97 samples described above, utilising either relative counts (as percentage of total sequencing reads for the sample) of 1) taxonomic data (genera), or 2) functional gene groups (COGs) to train RFs to predict a set of 11 environmental parameters. All parameters except NO_2 displayed a significant correlation ($P < 0.01$) between metagenome-predicted and measured parameter values (Figure 7; Table 1). Temperature, salinity and dissolved organic carbon (DOC) were the three parameters that were best predicted (Table 1). The most important genera for predicting the salinity level belonged to *Alphaproteobacteria*, *Gammaproteobacteria* and *Actinobacteria*, in accordance with these taxa gradually changing along the Baltic Sea salinity gradient.

	Taxonomic data		Functional data	
Parameter	**Pearson R**	**Spearman R**	**Pearson R**	**Spearman R**
Depth	0.69	0.49	0.75	0.5
Temp	0.94	0.93	0.95	0.93
Sal	0.94	0.89	0.94	0.82
NO2	0.12	0.43	0.08	0.44
NO3	0.66	0.58	0.69	0.59
NH4	0.62	0.64	0.38	0.5
PO4	0.81	0.59	0.76	0.54
SIO2	0.47	0.43	0.57	0.5
O2	0.63	0.62	0.57	0.55
Chla	0.64	0.66	0.63	0.7
DOC	0.76	0.75	0.81	0.72

Table 1. Correlations between metagenome-predicted and measured values for different environmental parameters. Predictions were made using counts of either taxonomic groups (genera) or functional gene groups (COGs; clusters of orthologous groups). Coefficients for Pearson and Spearman rank-order correlations are shown.

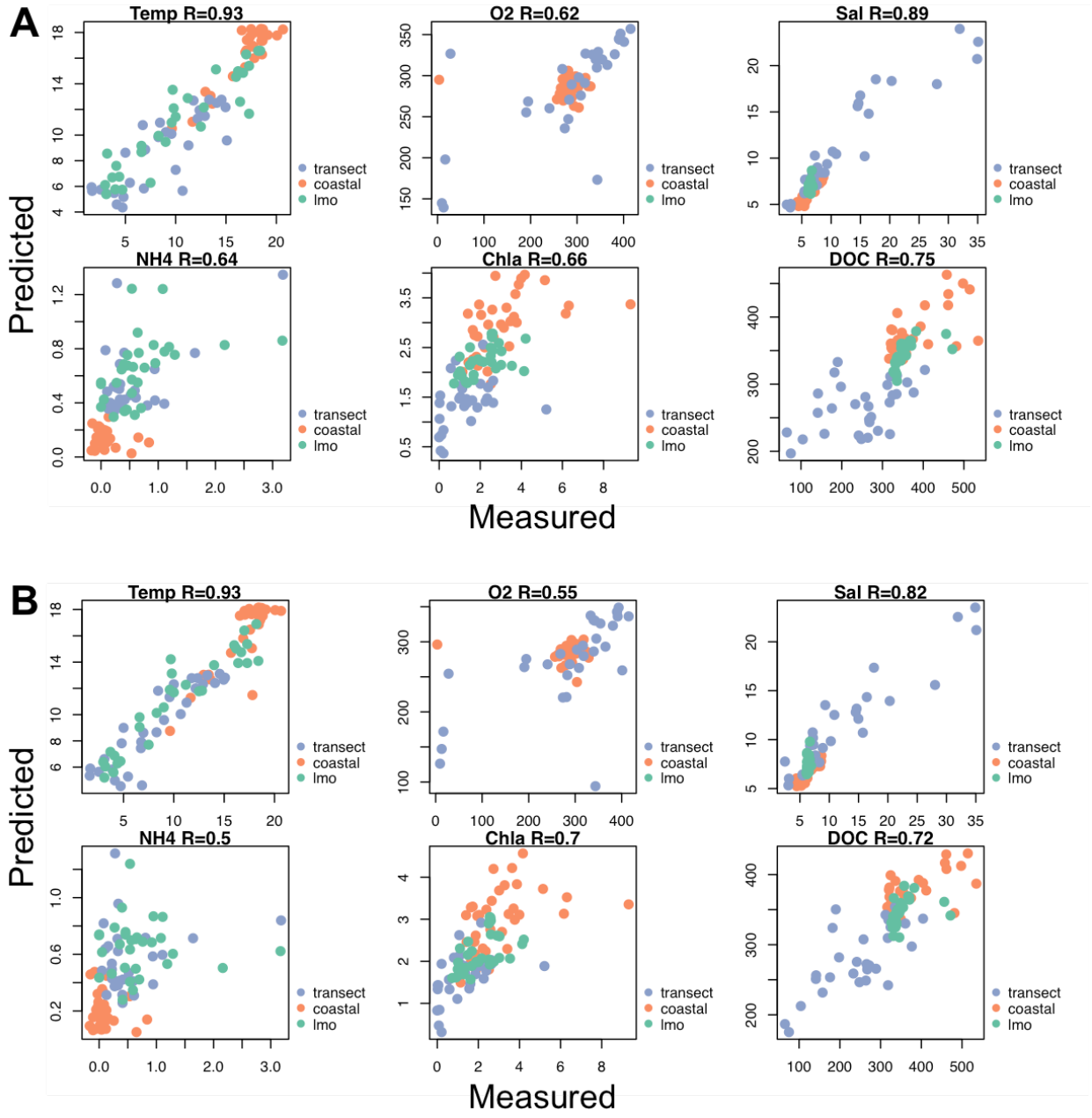

Figure 7. Measured (x-axes) and metagenome-predicted (y-axes) values for select parameters. Predictions were done using Random forests trained on either **A**) taxonomic (genera counts) or **B**) functional (COG counts) data from the metagenomes. The training and predictions were done using separate samples (out-of-bag predictions; Brieman 2001). Each dot is one sample, colored according to its sample group. Spearman rank-order correlation coefficients are indicated in the headers. Temp: Temperature (°C); O2 : O_2 (µmol/l); Sal: Salinity (PSU); NH4: NH_4 (µM); DOC: Dissolved organic carbon (µM); Chla: Chlorophyll *a* (µg/l).

Next we wanted to address how the predictions are influenced by sequencing depth (i.e. number of sequences read for each sample), since sequencing depth is correlated with cost. For the samples above, we obtained 3 - 103 million metagenome sequence reads per sample.

After taxonomic and functional annotation, 25,000 - 1.5 million and 360,000 - 22 million reads per sample were assigned to prokaryotic genera and gene functions (COGs), respectively. We subsampled different numbers of annotated reads from each sample before RF training and prediction. Figure 8 shows prediction accuracy as a function of number of annotated sequencing reads when predicting temperature in the samples. For gene-function data, prediction accuracy increase strongly with read depth, and work poorly at low depth, while for taxonomic data, already at 100 annotated reads (corresponding to on average 8000 starting reads) we get good predictions, and accuracy does not increase substantially with more reads.

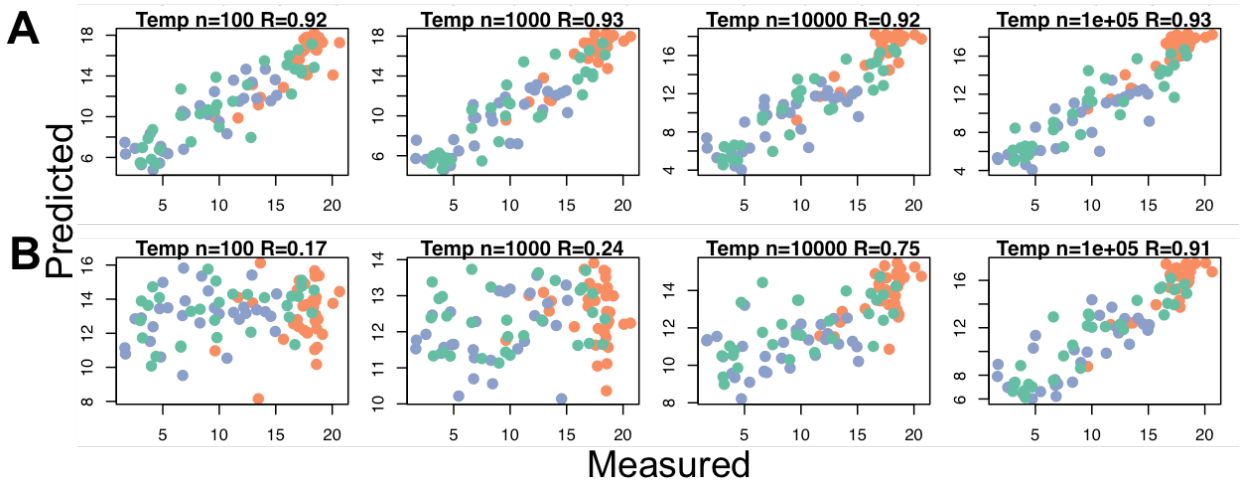

Figure 8. Measured (x-axes) and metagenome-predicted (y-axes) values for temperature (ºC) using varying number of annotated sequence reads using either **A**) taxonomic (genera counts) and **B**) functional (COG counts) data. Number of reads used is indicated above each plot.

Discussion and future perspectives

The BONUS Blueprint project is based on the prerequisite that a tight coupling exists between the bacterioplankton community and the physico-chemical properties of the water in which it resides. Before the onset of the project the scientific literature indicated that this would be the case (e.g. Herlemann et al., 2011), and this project has gathered significantly more evidence in support of this view. We show here using metagenome data how microbial communities change along the gradients of the Baltic Sea, both at an overall community level and for select genes and pathways.

Accuracy of predictions

The fact that we can predict environmental conditions from the metagenome data was one of the hopes of the project that had not been achieved before. Since then, a couple of studies have reported this type of analysis in aquatic systems. Firstly, Sunagawa et al. (2015) were able to predict temperature in the Tara Ocean global metagenome dataset with high accuracy (R = 0.86) and secondly Smith et al. (2015) used microbiome data to predict concentrations of a range of chemicals in groundwater, however with rather low accuracy. Predicting basic

parameters such as temperature, salinity or nutrient levels may appear to be of limited use since these can be measured by simpler means. However, if one can measure a whole spectrum of parameters using only the digital DNA read-out it may be both cheaper and quicker than using conventional methods for environmental monitoring. Further, the proof-of-principle presented here holds promise for a future wider use.

In our analysis we get similar accuracy of predictions when using taxonomic and functional classification of the metagenome data, but when using taxonomic data a smaller number of reads is needed. This is likely due to that the taxonomic data only included a few hundred different features (genera) while the functional data had several thousand features (COGs). The larger number of features for the functional data makes the abundance estimates for these features noisier due to fewer counts per feature. The fact that taxonomic data works well implies that, as an alternative to metagenomics, amplicon sequencing of taxonomic genes ('metabarcoding') can be used. This makes the analysis considerably cheaper since costs for sequencing library preparation and sequencing are an order of magnitude lower for metabarcoding (ca 30€ per sample) than for metagenomics (ca 300€ per sample). Moreover, bioinformatics analyses are considerably simpler and computational demands lower.

The added advantage with metagenomics compared to metabarcoding is, however, that it gives direct information on specific genes and pathways that may be of particular environmental or health relevance, and the data can be used to improve our understanding of the ecosystem and eventually create better ecosystem models.

Indicators for environmental status

An aim of the BONUS Blueprint project has been to explore the use of metagenomics data as a basis for indicators of environmental status in the Baltic Sea to assess progress in implementation of marine policies such as the HELCOM Baltic Sea Action Plan and the EU Marine Strategy Framework Directive (MSFD). Specifically we set out to identify a minimal set of genetic indicators that collectively, based on their relative abundances, are diagnostic for specific environmental conditions. The fact that basic parameters can be predicted with fairly high accuracy substantiates that metagenomics can be used as a basis for assessing environmental status and also indicates that we can train classifiers to predict other qualities of the water body that are harder to measure by conventional methods.

With the Random forest approach used here we are not restricting the analysis to a specific subset of functional genes or taxa, but instead the algorithm learns to associate patterns in the ensemble of features with environmental conditions. If metagenomic sequencing will be used for the future monitoring, there is no need to restrict the analysis to a smaller *a priori* defined set of features. However, if alternative, directed approached are to be used, such as hybridisation to gene-specific probes on gene chips, or quantitative PCR on panels of genes, it is of value to find specific genes of particularly high predictive power. To this end we identified the 10 most important functional gene families (COGs) for predicting each of the environmental conditions and trained the Random forest using the combination of these COGs. The results were as good as using the whole set of functional genes (data not shown) indicating that a substantially

smaller set of genes can be used for predicting the environmental conditions, opening up the possibility to use targeted approaches.

An example of when metagenomic indicators could be an advantage is in the screening for eutrophication in coastal areas as a complement to methods based on the more time-consuming establishment of inventories of macroalgae and zoobentos. The metagenomic data can also be used to assess genes indicative of specific pressures such as cyanobacterial toxins as shown here. Not least, indicators or indexes reflecting the diversity of the microbial community can be developed. Diversity indicators are specifically requested by the European Commission Decision on Good Environmental Status (EU 2017/848) to asses the impact of human activites on marine food webs. Using either metagenomics or metabarcoding, diversity indices for either the whole microbial plankton community or for specific taxonomic groups, such as different types of phytoplankton, can easily be obtained (e.g. Hu et al., 2016).

A challenge for the development of indicators based on metagenomics data is to define a threshold value that sets the boundary between ‘good’ and ‘not good’ environmental status. For existing indicators used in HELCOM or under the MSFD, this is typically done based on the use of long-term data sets to define a ‘reference period’ or by use of ‘reference sites’ with limited impact of human activities. With only a few years of metagenomic data available, and no pelagic environment in the Baltic Sea being unaffected by human activites, none of these options to define threshold values are available for the indicator based on metagenomics. Instead, we explored manipulations experiment carried out in the project to define a metagenome signal that corresponds to a significant effect by environmental pressure, i.e. to defined ‘not good status’. However, the experimental data do not cover a sufficient range of environmental conditions to support such an approach.

Detecting changes in central biogeochemical processes

The BONUS Blueprint datasets include vast opportunities to follow changes in biogeochemical processes of both ecological and management interest. In this context, the reduction targets for inputs of nitrogen and phosphorus to the Baltic Sea that have been agreed by the Baltic Sea countries can be highlighted. These reduction targets only make up a tenth of the measured turnover of nitrogen and phosphorus by microbes and thus, even minor changes to microbial processes may affect the results of management measures. In this report the results from a few genes and pathways involved in nitrogen and phosphorus cycling have been higlighted. With regard to nitrogen transformation pathways, we demonstrate that for genes involved in “opposing” processes such an nitrogen fixation and denitrification the spatial distribution pattern of the genes follows the expected patterns given what is known about the nitrogen cycle in the Baltic Sea. Although we did not measure biogeochemical process rates for these samples, the data indicates that the metagenome can be used to infer biogeochemically relevant transformations and the ratio between the central genes could tentatively serve as a proxy for directional changes in these central processes. Furthermore, the metagenomic datasets show the presence of genes involved in two processes that have received limited attention in the Baltic Sea so far; the potential for dissimilatory nitrate reduction (DNRA) and uptake of phosphonate. The substrate for DNRA is nitrite and the process is thus “competing” with

denitrification under similar environmental conditions. Phosphonate is typically not considered a source of phosphorus for microbial growth in the Baltic Sea, but in the BONUS Blueprint project is has been found that *Nodularia spumigena* carried a phosphonate degrading (*phn*) gene cluster that is activitated during experimental conditions (Teikari et al. 2018).

Monitoring

The bacterioplankton metatranscriptome, i.e. the expressed bacterial genes, would in theory be intimately linked to the environmental status as bacteria respond rapidly to contemporary environmental conditions. We planned to also evaluate metatranscriptome data for predictions, but due to the high cost we were not able to conduct metatranscriptomics on a large enough number of samples to allow comparison with the metagenome data. Our metatranscriptome data, and those of others (e.g. Aylward et al. 2015), display short-term temporal changes in gene expression, like a diurnal cycle. This extensive short-term variability cannot be reconciled with the frequency of sampling in marine monitoring programs; i.e. weekly or monthly sample acquisition. Moreover, sampling for and work with RNA is challenging. Consequently, for reasons related to expression variability, cost, and handling difficulty, we find that DNA based analyses of microbial plankton is more relevant for monitoring purposes.

As technology develops it also opens up new possibilities for DNA-based monitoring. Already today, cheap DNA sequencing machines are available that are as small as cell phones. It is reasonable to imagine that in a not-so-distant future buoys can be equipped with devices that autonomically samples and sequences DNA from the water and continously transfers sequences by satellite, enabling real-time monitoring of microbial taxa and encoded gene functions in local plankton communities. The intimate linkage between such data and environmental conditions demonstrated through the BONUS Blueprint project, and this report specifically, points to a potential high value of DNA sequencing methodology if applied in future marine plankton minotoring.

Acknowledgements

This work resulted from the BONUS Blueprint project supported by BONUS (Art 185), funded jointly by the EU and the Danish Council for Independent Research, Swedish Research Council FORMAS, Academy of Finland, Forschungszentrum Jülich GmbH (Germany), and the Estonian Research Council.

References

Alneberg J, Sundh J, Bennke C, Beier S, Lundin D, Hugerth LW, Pinhassi J, Kisand V, Riemann L, Jürgens K, Labrenz M, Andersson AF. 2018. “BARM and BalticMicrobeDB, a reference

metagenome and interface to meta-omic data for the Baltic Sea." *Scientific Data*. Accepted for publication (2018).

Aylward, Frank O., John M. Eppley, Jason M. Smith, Francisco P. Chavez, Christopher A. Scholin, and Edward F. DeLong. 2015. "Microbial Community Transcriptional Networks Are Conserved in Three Domains at Ocean Basin Scales." *Proceedings of the National Academy of Sciences of the United States of America* 112 (17):5443–48.

Bonaglia, Stefano, Isabell Klawonn, Loreto De Brabandere, Barbara Deutsch, Bo Thamdrup, and Volker Brüchert. 2016. "Denitrification and DNRA at the Baltic Sea Oxic-Anoxic Interface: Substrate Spectrum and Kinetics." *Limnology and Oceanography* 61 (5):1900–1915.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1). Kluwer Academic Publishers:5–32.

Dupont, Chris L., John Larsson, Shibu Yooseph, Karolina Ininbergs, Johannes Goll, Johannes Asplund-Samuelsson, John P. McCrow, et al. 2014. "Functional Tradeoffs Underpin Salinity-Driven Divergence in Microbial Community Composition." *PloS One* 9 (2):e89549.

Herlemann, Daniel Pr, Matthias Labrenz, Klaus Jürgens, Stefan Bertilsson, Joanna J. Waniek, and Anders F. Andersson. 2011. "Transitions in Bacterial Communities along the 2000 Km Salinity Gradient of the Baltic Sea." *The ISME Journal* 5 (10):1571–79.

Herlemann, Daniel P. R., Daniel Lundin, Anders F. Andersson, Matthias Labrenz, and Klaus Jürgens. 2016. "Phylogenetic Signals of Salinity and Season in Bacterial Community Composition Across the Salinity Gradient of the Baltic Sea." *Frontiers in Microbiology* 7 (November):1883.

Hugerth, Luisa W., John Larsson, Johannes Alneberg, Markus V. Lindh, Catherine Legrand, Jarone Pinhassi, and Anders F. Andersson. 2015. "Metagenome-Assembled Genomes Uncover a Global Brackish Microbiome." *Genome Biology* 16 (1):279.

Hu, Yue O. O., Bengt Karlson, Sophie Charvet, and Anders F. Andersson. 2016. "Diversity of Pico- to Mesoplankton along the 2000 Km Salinity Gradient of the Baltic Sea." *Frontiers in Microbiology* 7 (May):679.

Koskenniemi, Kerttu, Christina Lyra, Pirjo Rajaniemi-Wacklin, Jouni Jokela, and Kaarina Sivonen. 2007. "Quantitative Real-Time PCR Detection of Toxic Nodularia Cyanobacteria in the Baltic Sea." *Applied and Environmental Microbiology* 73 (7):2173–79.

Lindh, Markus V., Johanna Sjöstedt, Anders F. Andersson, Federico Baltar, Luisa W. Hugerth, Daniel Lundin, Saraladevi Muthusamy, Catherine Legrand, and Jarone Pinhassi. 2015. "Disentangling Seasonal Bacterioplankton Population Dynamics by High-Frequency Sampling." *Environmental Microbiology* 17 (7):2459–76.

Sivonen, K., K. Kononen, W. W. Carmichael, A. M. Dahlem, K. L. Rinehart, J. Kiviranta, and S. I. Niemela. 1989. "Occurrence of the Hepatotoxic Cyanobacterium Nodularia Spumigena in the Baltic Sea and Structure of the Toxin." *Applied and Environmental Microbiology* 55 (8):1990–95.

Smith, Mark B., Andrea M. Rocha, Chris S. Smillie, Scott W. Olesen, Charles Paradis, Liyou Wu, James H. Campbell, et al. 2015. "Natural Bacterial Communities Serve as Quantitative Geochemical Biosensors." *mBio* 6 (3):e00326–15.

Sunagawa, Shinichi, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, et al. 2015. "Ocean Plankton. Structure and Function of the Global Ocean Microbiome." *Science* 348 (6237):1261359.

Teikari, Jonna E., David P. Fewer, Rashmi Shrestha, Shengwei Hou, Niina Leikoski, Minna Mäkelä, Asko Simojoki, Wolfgang R. Hess, and Kaarina Sivonen. 2018. "Strains of the Toxic and Bloom-Forming Nodularia Spumigena (cyanobacteria) Can Degrade Methylphosphonate and Release Methane." The ISME Journal, February. https://doi.org/10.1038/s41396-018-0056-6.