The ISME Journal (2015) 9, 166–179

www.nature.com/ismej

OPEN

ORIGINAL ARTICLE

Satellite remote sensing data can be used to model marine microbial metabolite turnover

Peter E Larsen[1], Nicole Scott[2], Anton F Post[3], Dawn Field[4], Rob Knight[5], Yuki Hamada[6] and Jack A Gilbert[2,7]

[1]*Argonne National Laboratory, Biosciences Division, Argonne, IL, USA;* [2]*Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA;* [3]*The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA, USA;* [4]*NERC Centre for Ecology and Hydrology, Wallingford, UK;* [5]*Department of Chemistry and Biochemistry, BioFrontiers Institute, University of Colorado at Boulder, Boulder, CO, USA;* [6]*Argonne National Laboratory, Environmental Science Division, Argonne, IL, USA and* [7]*Argonne National Laboratory, Institute for Genomic and Systems Biology, Argonne, IL, USA*

Sampling ecosystems, even at a local scale, at the temporal and spatial resolution necessary to capture natural variability in microbial communities are prohibitively expensive. We extrapolated marine surface microbial community structure and metabolic potential from 72 16S rRNA amplicon and 8 metagenomic observations using remotely sensed environmental parameters to create a system-scale model of marine microbial metabolism for 5904 grid cells (49 km^2) in the Western English Chanel, across 3 years of weekly averages. Thirteen environmental variables predicted the relative abundance of 24 bacterial Orders and 1715 unique enzyme-encoding genes that encode turnover of 2893 metabolites. The genes' predicted relative abundance was highly correlated (Pearson Correlation 0.72, *P*-value $<10^{-6}$) with their observed relative abundance in sequenced metagenomes. Predictions of the relative turnover (synthesis or consumption) of CO_2 were significantly correlated with observed surface CO_2 fugacity. The spatial and temporal variation in the predicted relative abundances of genes coding for cyanase, carbon monoxide and malate dehydrogenase were investigated along with the predicted inter-annual variation in relative consumption or production of ~3000 metabolites forming six significant temporal clusters. These spatiotemporal distributions could possibly be explained by the co-occurrence of anaerobic and aerobic metabolisms associated with localized plankton blooms or sediment resuspension, which facilitate the presence of anaerobic micro-niches. This predictive model provides a general framework for focusing future sampling and experimental design to relate biogeochemical turnover to microbial ecology.

The ISME Journal (2015) **9,** 166–179; doi:10.1038/ismej.2014.107; published online 29 July 2014

Introduction

The oceans comprise 72% of the planet's surface area, and harbor microbial communities responsible for 98% of the ocean's primary productivity (Jørgensen and Boetius, 2007). Efforts to characterize the microbial ecology of the oceans have always relied on local longitudinal (Gilbert *et al.*, 2012) or geospatially distributed (Fuhrman *et al.,* 2008) observations to make inferences about the ecological dynamics of microbial community structure and function as a product of local niche variability. Such inferences make extrapolative assumptions regarding the variation in niche dynamics both in space and time, which is essential as mapping the real-time response of microbial communities across the global ocean would be prohibitively expensive. Researchers have therefore chosen to model the microbial dynamics of marine ecosystems to develop a predictive understanding of the microbial response to environmental change. However, capturing these dynamics has focused on predicting the relative abundance of key taxonomic groups, with only limited characterization of functional capacity (Follows and Dutkiewicz, 2011; Larsen *et al.,* 2012; Ladau *et al.,* 2013; Toseland *et al.,* 2013; Fierer *et al.,* 2013). Going beyond predictions of the membership of microbial communities, to predict their functions, has been an elusive goal. However, physical, chemical and biological data collected in the Western English Channel (WEC) provide a unique opportunity to generate and validate microbial community distribution models (Fierer and Ladau, 2012)

Correspondence: JA Gilbert, Institute of Genomics and Systems Biology, Argonne National Laboratory, 9700 S Cass Avenue, Argonne, IL 60439, USA.
E-mail: gilbertjack@anl.gov
Received 18 February 2014; revised 25 April 2014; accepted 28 May 2014; published online 29 July 2014

that can predict microbial metabolic potential. Station 'L4' of the Western Channel Observatory is an oceanographic time-series and marine biodiversity reference site (http://www.westernchannelobservatory.org.uk) that contains a rich resource of oceanographic, climatological, remote sensing, biogeochemical and biological data, including molecular characterization of the microbial community with 16S rRNA (bacteria and archaea) and shotgun metagenomic and metatranscriptomic sequence data (Gilbert *et al.,* 2009, 2010, 2012; Caporaso *et al.,* 2012; Gibbons *et al.,* 2013). We combined this community sequence data resource with synoptic environmental parameter data predicted from remote sensing satellite image models (SIMs) covering the whole English Channel (http://ncof.npm.ac.uk/), an indispensable tool for assessing environmental conditions over wide spatial and temporal scales in macroecological systems (Graetz, 1990; Paul Bissett *et al.,* 2001; Brewin *et al.,* 2010; Wang *et al.,* 2010) and microbial systems (for example, Doney *et al.,* 2004; Glöckner and Joint, 2010). Remote sensing measures reflected or emitted electromagnetic radiation without requiring a physical presence in the region being monitored; *in situ* environmental conditions can then be extrapolated by correlation with shifts in the spectral reflectance using SIMs. Because remote-sensing data can be captured inexpensively at large scales and relatively fine spatio-temporal resolution, and is already being collected by various national and international agencies (for example, National Aeronautics and Space Administration and the European Space Agency), it represents a valuable resource for extrapolating ecosystem dynamics. The unique combination of available sequence data describing the microbial community and SIM data at the WEC site has enabled the development of two tools for creating community distribution models: Microbial Assemblage Prediction (MAP; Larsen *et al.,* 2012) and Predicted Relative Metabolic Turnover (PRMT; Larsen *et al.,* 2011). MAP predicts microbial community structure as a function of SIM data by creating a Bayesian co-dependency network where the SIM data are parents to microbial taxa, and microbial taxa can also be parents to other microbial taxa; this artificial neural network was trained on longitudinal data describing microbial community structure from the L4 site (Gilbert *et al.,* 2012) and these predicted community structures agreed well with the actual community structures (Bray–Curtis dissimilarity 0.897; Larsen *et al.,* 2012). This model forms the basis of the current study. PRMT is a translation tool that uses the changing relative abundance of functional genes in metagenomic data, or functional transcripts in metatranscriptomic data, between samples to predict the changing capacity of that community to consume or generate metabolites, for example, CO_2, NO_3, NH_4, CH_4 and so on (for example, Larsen *et al.,* 2011; Scott *et al.,* 2014; Mason *et al.,* 2014). Negative PRMT scores indicate increased relative metabolite production and hence accumulation, and positive scores indicate relative metabolite consumption. PRMT does not predict fluxes of metabolites or gases, nor does it provide a direct quantitative approach for measuring the concentration of metabolites in a sample. It does, however, translate the metabolic potential of a microbial assemblage by utilizing the relative abundances of all annotated genes encoding potential enzyme functions that influence the consumption or production of metabolites, to improve an estimation of the relative likelihood that a metabolite will be consumed or accumulated by that assemblage compared with another assemblage. PRMT was previously validated against measured concentrations of marine metabolites from the L4 site, demonstrating that it was possible to predict the changing capacity of the community to mediate the transformation of those metabolites from the changing relative abundance of gene fragments that code for enzymes in metagenomic data (Larsen *et al.,* 2011).

Here we combine MAP and PRMT approaches to generate a system-scale model of marine microbial metabolism as a function of environmental properties that were estimated from remotely sensed spectral reflectance data. SIM data are used to predict community structure with MAP, and community structure is used to predict the community metagenome (as the relative abundance of unique enzyme activities, using a technique similar to PiCRUST (Langille *et al.,* 2013), which is translated, using PRMT, into the relative capacity of the community to consume or generate different metabolites; Figure 1). Although it is impossible to absolutely validate these models, as data simply does not exist, we have correlated these statistical models against observed amplicon and metagenomic shotgun sequencing data collected from the L4 station (Gilbert *et al.,* 2010, 2012). In addition, although PRMT scores are only an indication of the relative potential of the predicted assemblages to turnover metabolites, we show that the PRMT scores for CO_2 correlate appropriately with 1700 *in situ* data points for the fugacity of CO_2 (fugacity is an effective measure of the pressure of a gas that replaces real mechanical pressure in equilibrium equations) acquired from the Surface Ocean CO_2 Atlas (SOCAT; http://www.socat.info). We discuss how these measurements can be used to make informed predictions about the distribution and temporal dynamics of functional metabolisms in this ecosystem.

Materials and methods

Data used to train the MAP-PRMT model

L4 Station amplicon and metagenomic sequence data. The L4 Station of the Western Channel Observatory is an oceanographic time-series and marine biodiversity reference site (http://www.westernchannelobservatory.org.uk) that provides a

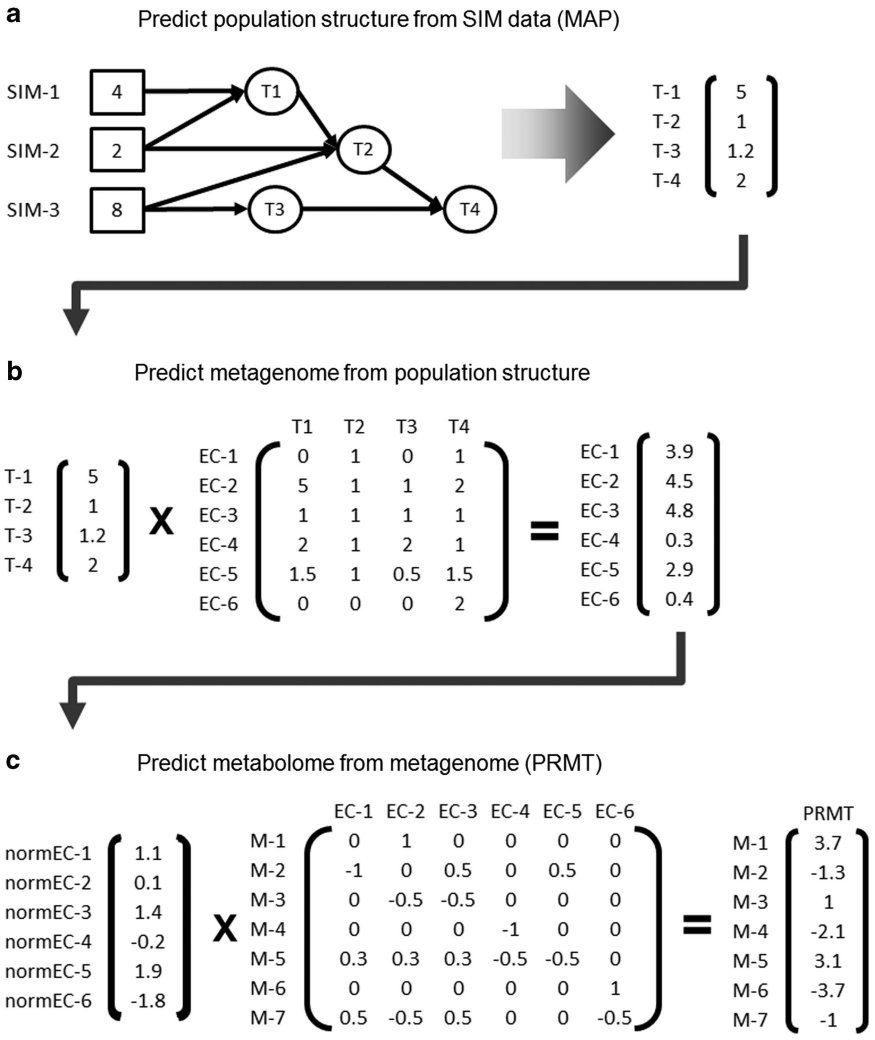

Figure 1 The procedure for modeling eco-scale metabolome from SIM data in the WEC follows three principle steps, outline here in cartoon form. In the first step (**a**), SIM data (labeled SIM-1 to 3) are used to predict the community structure of microbial taxa (labeled T-1 to 4) using an artificial neural network trained on SIM data and 16S rRNA amplicon data. The output of this step is a vector describing the relative abundance of taxa for a given set of SIM data. The output of step (**a**) is used as the input for step (**b**). The vector of microbial community structure is transformed into a predicted metagenome. Each element in the matrix in **b** is the average number of genes for an enzyme activity (labeled EC-1 to 6) in each microbial taxa T. The output is a predicted metagenome described as a vector of enzyme function counts. The output from **b** is normalized by the average enzyme function counts for all predicted metagenomes and used as input for step (**c**). In **c**, metagenome is used to predict the metabolic turnover for metabolites mediated by the enzyme functions in **b**. The matrix in **c** is a metabolomic network derived from enzyme functions in metagenome and the set of metabolites (labeled M-1 to 7) inferred by those enzymes. The output of step (**c**) is a vector of PRMT-scores (Predicted Relative Metabolic Turnover), a quantification metric for predicted relative turnover of every metabolite in the metabolomic network. Steps **a**–**c** are performed for every location and every time point in the WEC for which SIM data are available.

rich collection of bacterial and archaeal 16S rRNA, and shotgun metagenomic and metatranscriptomic sequence data sets over seasonal cycles (Gilbert *et al.,* 2010, 2012). The average water depth for the English Channel is 125 m, and remains relatively well mixed throughout the year but with defined seasonal stratification (Smyth *et al.,* 2009). These amplicon and metagenomic sequence data describe the taxonomic and functional dynamics of the microbial communities and are abundantly contextualized with environmental parameters, including *in situ* measured physical, biological and chemical variables (Southward *et al.,* 2005; Smyth *et al.,* 2009). Eight samples taken at the L4 WEC station over Spring, Summer and Winter 2008 were processed to generate shotgun metagenomic sequence data (Gilbert *et al.,* 2010). These were used as a baseline to extrapolate microbial functional potential across time and space in the English Channel and validate predictions of the relative abundance of metagenomic sequences predicted to code for proteins with annotated enzyme activities. To facilitate the predictions of population and functional structure, the WEC region is divided into 5904

(123×48) grid cells (Larsen *et al.,* 2012), and predictions were generated for weekly averages from 2007 to present day.

Remote sensing SIM data. SIM data were used as generated for the WEC Ecosystem model (ECOOP; WMS 1.3.0 data from National Centre for Ocean Forecasting GODIVA). These data are derived from moderate-to-high temporal frequency coarse spatial resolution satellite data such as Moderate Resolution Imaging Spectrometer (daily–weekly; 250 m–1 km resolution) and Advanced Very High Resolution Radiometer (daily; 1∼4 km resolution). These time-series satellite data were correlated with ocean environmental parameters (for example, sea surface temperature, chlorophyll and phosphate concentrations, and dissolved oxygen) measured at the sampling locations using rigorous algorithms and models (Kilpatrick *et al.,* 2001; Carder *et al.,* 2004) to generate spatially explicit and contiguous information. The remotely sensed environmental information is available for the L4 Station (http://ncof.npm.ac.uk/ncWMS/godiva2.html; Southward *et al.,* 2005; Smyth *et al.,* 2009) for dates between 2007 and April 2010. Thirteen SIM data types (dissolved O_2, phosphate, nitrate, ammonium, silicate, chlorophyll A, photosynthetically active radiation (PAR) irradiance, small particulate organic carbon (POC), medium POC, large POC, labile dissolved organic carbon (DOC), semi-labile POC and bacteria) were used in the MAP model for the current study. SIM data are reported for 1200 hours each day.

SOCAT. The SOCAT is the database of a global ocean fCO2 (fugacity of CO_2) measurements presented in a common format (Pfeil *et al.,* 2013). Fugacity is a relative measure of gas pressure, which replaces mechanical pressure in equilibrium equations; this is the metric used to represent the pressure of CO_2 in the water column samples from the SOCAT database. There were 1798 points for the year 2008 in the SOCAT database of marine fCO_2 that overlapped in time and space with the previously generated English Channel microbial community distribution model predictions and available SIM data (Larsen *et al.,* 2012). Of these data points, we used 244 collected between 1100 hours and 1300 hours, corresponding to the approximate time of SIM data collection (1200 hours) to correlate predicted PRMT scores for CO_2 against *in situ* measured fCO_2.

Prediction of microbial community structure

The MAP model (Larsen *et al.,* 2012) extrapolates community structure (relative abundance of defined taxonomic units, for example, bacterial Orders) as a functional of SIM data. MAP requires a set of sampled microbial assemblages (16S rRNA amplicon sequence data) and corresponding measures of environmental parameters (for example, SIM data; Figure 1a). These data are used to generate a directed acyclic graph, called the Environmental Interaction Network (EIN). Nodes in the EIN are environmental parameters and measures of microbial taxonomic relative abundances. Edges in the EIN represent potentially causal relationships between environmental parameters and microbial taxa, or between different microbial taxa. Root nodes in the EIN are exclusively comprised of the environmental parameters. We use the previously developed network created using Bayesian network prediction methods (BANJO v 2.0.1 (Yu *et al.,* 2004; Smith *et al.,* 2005)) from the MAP model (Larsen *et al.,* 2012). Let p_i^x be the proportion of a single taxa i in sample x. The community structure for a sample is given by an array of length T:

$$\overrightarrow{p^x} = \{p_1^x, p_2^x \ \ldots \ p_T^x\} \quad (1)$$

where T is number of represented taxa. Each $\overrightarrow{p^x}$ is normalized such that the sum of community abundance values in a single sample equals 100. The value of each taxa (where each taxon is a node in the EIN) in a sample, p_i^x, is predicted as a function of the values of the m parents of the ith taxa in the EIN network:

$$p_i^x = f(p_{i,1}^x, p_{i,2}^x \ \ldots \ p_{i,m}^x) \quad (2)$$

Parents of a given taxon may be other taxa or environmental parameters. As the EIN is a directed acyclic graph with root nodes representing environmental parameters, all predicted values for taxonomic abundances are ultimately functions of environmental parameters. The functions to predict the taxa values are generated using an evolutionary optimization algorithm tool EUREQA (v 0.83 beta software (Schmidt and Lipson, 2009)). Here we use the previously generated MAP functions (Equation 2) to predict the relative abundance of the 24 most abundant bacterial Orders in the WEC from 24 monthly measurements of 16S rRNA V6 community structure corresponding to 2007–2008 of the 6-year time-series from L4 (Gilbert *et al.,* 2012). This model was generated using 16S RNA gene sequence observations collected over a 2-year period (2007–2008) at the L4 Station (Larsen *et al.,* 2012) and 13 environmental parameters calculated from SIM data (including dissolved O_2, phosphate, nitrate, ammonium, silicate, chlorophyll A, photosynthetically active radiation irradiance, small POC, medium POC, large POC, labile dissolved organic carbon, semi-labile POC and bacteria). As previously reported, the predicted microbial community structures matched observed community structures well (average Bray–Curtis similarity score of 89.7 (s.d. = 2.32) over the 24 monthly observations).

Prediction of enzyme relative abundance

The community structure of a sample can also be used to predict the community functions of that population, given as relative abundances of specific

functional gene annotations represented as Enzyme Commission (EC) numbers. This involves collation of all available genomes from a given taxon group (for example, Order level) to enable an averaged predicted metagenome for a given community (Figure 1b). Let g_i^x be the abundance of a single-enzyme function i in sample x. The functional structure for a sample is given by an array of length EC:

$$\overrightarrow{g^x} = \{g_1^x, g_2^x \ldots g_{EC}^x\} \quad (3)$$

Where EC is the number of represented EC number annotations for unique enzyme activities. To predict the functional structure $\overrightarrow{g^x}$, the population structure of the sample p_i^x is 'weighted' by the average distribution of genes coding for specific enzyme annotation for all annotated genomes belonging to taxa t. Let the enzyme by taxa matrix be, E:

$$\overrightarrow{g^x} = E\,\overrightarrow{p^x} \quad (4)$$

We calculate E by using a database, such as the collected annotated genomes in kyoto encyclopedia of genes and genomes (KEGG), to determine the average EC number counts for each taxa present in the community. In the current study using the 24 bacterial taxa in the previously published MAP model, we used a total of 1398 representative sequenced and annotated genomes present in the KEGG database to determine E (Supplementary Data: Predicted average abundance of EC annotations per bacterial Order (Supplementary Table 1); Complete listing of KEGG genomes used (Supplementary Table 2)). All g_i^x are normalized by quantiles across all x samples and $\log_2$ transformed, $\overrightarrow{g'^x}$. The 1042 enzymes that overlap between the predicted and observed metagenomic data are shown in Supplementary Table 3.

Prediction of Microbially Mediated Relative Metabolite Turnover (PRMT)

PRMT is a method for quantifying the predicted difference in relative metabolic turnover between samples as a function of the abundances of genes predicted to code for enzyme activities in metagenomic sequence data (Larsen *et al.,* 2011). PRMT scripts are available at http://www.bio.anl.gov/PRMT.html. Let the enzyme by metabolic reaction matrix be *M*. *M* is of the dimensions EC by m, where m is the number of metabolites predicted from the given number of enzyme functions (Figure 1c). The metabolic turnover scores between samples, x and y, are given by:

$$\overrightarrow{c_{x,y}} = M(\overrightarrow{g'^x} - \overrightarrow{g'^y}) \quad (7)$$

We calculate M by using a database such as the KEGG reactions to determine the counts of compounds involved in reactions dictated by the enzymes present in the normalized and $\log_2$ transformed metagenomes, $\overrightarrow{g'^x}$ and $\overrightarrow{g'^y}$.

The resulting set of value, $\overrightarrow{c_{x,y}}$, is a vector of PRMT scores of length m for the comparison of PRMT of each metabolite in M for population x relative to population y.

We previously demonstrated that the PRMT analysis of metagenomes collected in the WEC strongly correlated with corresponding measured environmental parameters (Larsen *et al.,* 2011). Importantly, although there is a linear relationship between the relative abundance of genes encoding enzymes and PRMT scores for a given metabolite, there is also a nonlinear relationship between each of the genes encoding enzymes that may be used to predict a given metabolite. Therefore, the contribution of each enzyme is weighted based on the network topology, and a supposed equilibrium between production and consumption. All scripts used for PRMT are available via http://bio.anl.gov/prmt.html. All PRMT scores provided in this work are presented in Supplementary Material (Supplementary Table 4).

Validation of predicted relative abundances of enzyme activities and CO_2 relative turnover

To determine whether the predicted gene and enzyme activity relative abundances were representative of the observed metagenomes sequenced from the L4 site, eight metagenomes (relative abundances of functional genes) were predicted using the method outlined above, for the dates and times for which the observed metagenomes exist for the L4 site in the WEC (Gilbert *et al.,* 2010). As the method outlined above only predicts the relative abundance for the annotated enzyme activities associated with the taxonomic Orders that were generated by the MAP model, not all EC annotations present in the observed shotgun metagenomes were found in the predicted metagenome. In fact, only 1042 enzyme functions were in common between the predicted metagenomes and the observed metagenomes. However, a Pearson Correlation Coefficient (PCC) was used to determine the correlation between the predicted and observed relative abundances of the 1042 unique enzyme activities (as EC number annotations) found in common between these two data sets. Total enzyme function counts from both data sets were log transformed before calculating correlations.

We also validated the method by comparing the CO_2 metabolite turnover predictions generated from PRMT to measured metabolic parameters. Currently, the L4 site in the Western Channel Observatory is the most characterized location in the English Channel (Smyth *et al.,* 2009), and we previously used metabolite concentrations measured *in situ* at this site to validate PRMT, whereby we were able to predict changes in the concentrations of chlorophyll A (PCC −0.98), organic nitrogen (PCC −0.99),

organic carbon (PCC −0.98), nitrate (PCC −0.98), ammonia (PCC −0.81) and orthophosphate (PCC −0.93) to a high degree of accuracy (Larsen *et al.,* 2011). However, only one uniformly measured *in situ* metabolite concentration spanned the English Channel over multiple years: the fugacity of CO_2 in the SOCAT database. We used these reported values collected across the WEC region at a variety of dates. PRMT scores were simulated for locations and times that mapped the 1744 observed WEC SOCAT fCO_2 data points, of which 244 mapped to the times and locations for which the SIM data were generated; these were used for determining correlations between the calculated PRMT-CO_2 and measured fCO_2. The significance of all correlations was determined using a bootstrap approach. The Bray–Curtis score between the predicted and actual community structure was determined, and all predicted data were randomly permuted 10 000 times (randomizing both the sample day and taxa). Then, for each permutation, the Bray–Curtis score was recalculated. Each time a random permutation of data returned a Bray–Curtis value higher than the initial Bray–Curtis value it was recorded, and a *P*-value was calculated. The seasonal variance over metabolites with non-zero PRMT scores (2494 out of 2893 possible metabolites) over 3.5 years of weekly averages of the channel were grouped using K-means clustering. Clustering was performed in Multi Experiment Viewer (MeV) v4.5.1 (http://www.tm4.org/) for six clusters, using Euclidian distance. Clustering was run iteratively until convergence, with no clusters taking longer than eight iterations. The six resulting clusters followed patterns of seasonal variation, with cluster (i) (108 metabolites) with high PRMT scores in summer, cluster (ii) (114 metabolites) up in winter, cluster (iii) (103 metabolites) dipping sharply in winter, cluster (iv) (163 metabolites) spiking in winter, cluster (v) (1189 metabolites) slightly up in winter and cluster (vi) (817 metabolites) slightly up in summer.

Results and discussion

Validation of microbial community structure and annotated enzyme function predictions

We demonstrate that satellite data describing surface water reflectance properties can be used to predict not only the microbial community structure (as previously shown; Larsen *et al.,* 2012), but also the relative abundances of annotated genes that encode enzyme activities, which can then be translated into the changing capacity of the community to consume or generate metabolites. As demonstrated previously, the MAP model can be used to predict, with significant accuracy, the community structure (relative abundances of the 24 most abundant bacterial Orders) in a longitudinal data set to a Bray–Curtis dissimilarity score of 0.897 (Larsen *et al.,* 2012). We now extend this work by predicting the relative abundances of genes in these communities, and validate these predictions by comparisons to sequenced metagenomes. A significant correlation of 0.718 (P-value $<10^{-6}$) was observed when comparing the relative abundances of the 1042 annotated enzyme activities (Supplementary Table 3) for eight predicted metagenomes from this study that overlap in time and space with eight observed metagenomes from a previous study (Gilbert *et al.,* 2010). This provides confidence that the predicted community structure and functional potential generated for the 5904 sites in the WEC significantly represented observed structure. This method is limited to genes that can be annotated to known enzyme activities, and the attribution of these functions to known taxa is limited to the level of bacterial Order. However, the framework can easily be extended when new gene annotations and genomes become available. Although predicting species level descriptions, that is, not generalizing functional predictions to the level of bacterial Order, would be extremely useful, genomes that effectively describe the species level functional potential are lacking for many taxa, which seriously impacts the effectiveness of species level predictions. It is more effective to generalize at a higher taxonomic level (Langille *et al.,* 2013). Also extrapolating over such wide geographic and temporal ranges necessitates a certain degree of taxonomic generalization, to reduce the likelihood of over fitting local genomic adaptations. That the predicted relative abundance of annotated enzyme-encoding genes used to generate PRMT scores was significantly correlated with observed metagenomic data, builds on previous evidence that a metagenome can be predicted by correlating functional gene abundance to microbial community structure as measured by 16S rRNA amplicons (Langille *et al.,* 2013).

Comparison of metabolic predictions with observed fCO_2

Although the model can be used to accurately predict the microbial community structure and the relative abundance of functional genes at a particular site, its true value is in the potential to extrapolate and generate modeled hypotheses regarding predictions of metabolite turnover, that is, the emergent metabolic properties of the microbial assemblage for a given location and time as a function of environmental niche space (for example, predicted from SIM data). To determine whether extrapolations of the relative turnover of metabolites reflected what was observed across the channel, we compared the results against *in situ* metabolite measurements from multiple times and sites. However, such data were only available for one metabolic measurement, the fugacity of CO_2 (fCO_2), as provided by the SOCAT database. Therefore, we predicted the relative turnover of CO_2 for 244 locations and time points in the

English Channel where available SIM data and *in situ* fCO_2 measurements overlapped. Enzymes that mediate the bidirectional conversion of metabolites to CO_2 do not directly contribute to the predicted PRMT scores for CO_2; therefore, unidirectional enzymes have greater significance. In this calculation, genes encoding 1.4.4.2, 2.3.1.-, 2.7.2.2 and 3.5.1.54 (Supplementary Figure 2) provide the greatest contribution to CO_2 accumulation. Comparing PRMT scores for CO_2 with 244 observed fCO_2 data points provided a significant correlation of -0.4 ($P<0.0001$). Therefore, when the model predicted an increased relative capacity for the synthesis, and hence accumulation of CO_2 (negative PRMT score), the observed fCO_2 was greater than the average across the 5904 grid cells for that time point, which suggested a local increase in the relative pressure of CO_2. Conversely, when the model predicted an increase in the relative capacity for the consumption of CO_2 (positive PRMT score), the observed fCO_2 was relatively lower, which suggested a decrease in the relative pressure of CO_2. This result demonstrates that the model is able to reflect dynamics in the metabolic turnover of CO_2 across the English Channel and over time, indicating that other predicted metabolite turnover scores may also be representative. We propose that these predictions could be used to help direct efficient sampling efforts to investigate specific microbial biogeochemical processes. Importantly, Although cyanobacteria and bacterial heterotrophs will impact the fugacity of CO_2, we do not know the degree of their influence. The moderate, but significant, correlation between predicted turnover and fCO_2 observed here may reflect a limited influence, which is probably less important than the physical equilibrium dynamic of CO_2 at the water–atmosphere interface. Therefore, despite this significance, these PRMT scores and this correlation should in no way be used to infer true fCO_2 dynamics, yet the PRMT scores of all metabolites (as outlined below) can be used to explore interesting extrapolated dynamics across a region and through time.

Extrapolating central carbon metabolism across the WEC

To highlight the model's potential to capture functional gene abundance and metabolite turnover features across the channel, and to demonstrate how the predicted relative abundances of genes relate to the PRMT scores for the metabolites they mediate, the changing relative abundance of three genes associated with CO_2 metabolism were predicted, and compared with the relative turnover of CO_2 (Figure 2). Predictions of microbial community structure (relative abundance of bacterial orders), functional gene abundance and PRMT scores were generated for 5904 grid cells (49 km^2) in weeks commencing 17 March, 23 June, 22 September and 8 December 2008, which were arbitrary chosen weekly averages for which adequate SIM data were available (Figure 2; Supplementary Figure 1). This resulted in a total of 566 784 predictions of the relative abundance of the 24 most abundant bacterial Orders; 40 501 440 functional gene relative abundance predictions; and 68 321 088 separate PRMT scores. We chose to demonstrate these predictions using three central carbon metabolism enzymes associated with CO_2 production and consumption: cyanase, carbon-monoxide (CO) dehydrogenase and malate dehydrogenase (Supplementary Figure 2). These three enzymes (or more accurately, the relative abundance of the genes that encode them) were chosen to represent three different types of CO_2 metabolism, which are differentially affected by environmental conditions, are representative of different microbial populations, and exemplify how these differences interact synergistically to affect predicted CO_2 turnover (PRMT score). Cyanase is involved in toxicity; CO dehydrogenase is considered an obligate anaerobic system of carbon fixation and is found in both bacteria and archaea (although it is also found in aerobic carboxydovores); and malate dehydrogenase is part of the reversible conversion of malate to oxaloacetate, and forms part of both the oxidative (tricarboxylic acid cycle, TCA) and reductive TCA (rTCA) in bacteria.

Strikingly, our model predicts a very similar distribution for the normalized copy number of genes encoding these pathways for the weekly averages in March and December (Figure 2). Although this could be an artifact, it may also indicate that the community functional potential for these genes is relatively similar in these 2 months, which have somewhat similar environmental characteristics, for example, low temperature and high inorganic nutrient concentrations (Smyth *et al.,* 2009). This would need to be validated by *in situ* observation, but such an endeavor is outside the scope of the presented work. The spatial variance in the predicted copy number of the gene encoding cyanase during June and September suggests spatial heterogeneity in cyanase metabolism increases during these months. The predicted copy number of the gene encoding carbon-monoxide dehydrogenase demonstrates high relative abundance along the north coast of France and the southern Irish Sea in June and September, which is replaced by relatively higher copy numbers in the mid channel during the March and December (Figure 2). The predicted normalized copy number of the gene encoding malate dehydrogenase shows the most homogeneous distribution. Cyanase (4.2.1.104) mediates the bi-directional bicarbonate-dependent degradation of cyanate to CO_2, to detoxify cyanate to carbamate, which spontaneously degrades to NH_3 and CO_2. The cynS gene in marine cyanobacteria is thought to have a role in detoxification of cyanate, which builds up becaue of urea and carbamoyl phosphate metabolism (Kamennaya and Post, 2011). Although difficult to test with these data, areas of increased cynS abundance may represent regions that

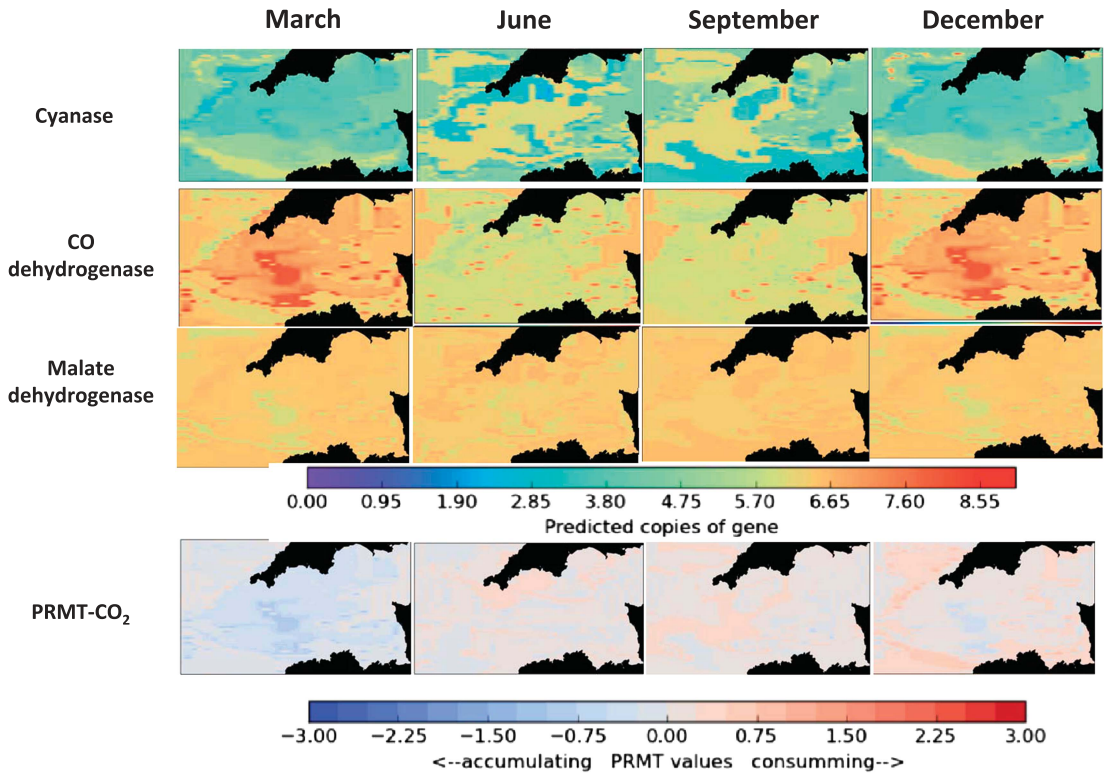

Figure 2 Maps of the Western English Channel showing the changing relative abundance of three central carbon metabolism genes and PRMT scores for carbon dioxide over 4 monthly averages in 2008. The three top rows show weekly averages for the relative gene copy number for Cyanase (4.2.1.104), Carbon-monoxide dehydrogenase (1.2.99.2) and Malate dehydrogenase (1.1.1.37). As the predicted metagenomes are created by merging the relative proportions of genomes from taxa closely related to the 16S rRNA taxa, the copy number refers to the number of copies of that gene found in that predicted metagenome, as these numbers are often very small we decided to show the absolute value rather than the relative percentage. These are 3 enzymes (out of 30) that contribute to the PRMT scores for CO_2 as calculated in the fourth row. For each graph, values were averaged across each day within a week during the given month. The predicted copies of a gene in each calculation are normalized, and therefore are not absolute. PRMT scores in red suggest a relative capacity of that assemblage to consume CO_2, whereas scores in blue suggest a relative capacity to accumulate CO_2.

require such de-toxicity, which is supported by its increased abundance across the channel during the months of high phytoplankton productivity (March and September). This is consistent with previous observations from various marine basins (Red Sea, Mediterranean Sea, Indian Ocean and Southern Ocean) that periods of high productivity induce labile dissolved organic matter and nitrogen (urea and amino acids), which can trigger cyanase production (Kamennaya and Post, 2013). Carbon-monoxide (CO) dehydrogenase (1.2.99.2) is central to the carbonyl branch of the reductive acetyl-CoA pathway, and is considered to be the most ancient autotrophic carbon fixation pathway becaue of its presence in both the bacteria and archaea (Hügler and Sievert, 2011). However, the genes associated with these predictions are almost exclusively bacterial in origin. CO dehydrogenase is mostly an obligate anaerobic system of carbon fixation, which would suggest a significant increase in anaerobic micro-niches associated with the phytoplankton and zoo-plankton blooms occurring mid-channel during the summer months. Although aerobic CO oxidation can also be indicated by this gene, it is limited to specific organisms belonging to the Actinobacteria, Proteobacteria and Firmicutes (Martin-Cuadrado *et al.,* 2009), and in this model CO dehydrogenase is mostly associated with Desulfobacterales and Sphingomonadales, which to the best of our knowl-edge do not comprise carboxydovores. This suggests that the majority of predicted CO dehydrogenase activity may be anaerobic. Malate dehydrogenase (1.1.1.37) is primarily responsible for the reversible redox of malate to oxaloacetate, and forms part of both the oxidative (TCA) and rTCA in bacteria and the dicarboxylate/4-hydroxybutyrate cycle in archaea. Therefore, this enzyme occurs in both aerobic and anaerobic systems, and indeed could represent both pathways in the predictions of CO_2 turnover. However, it is more likely that the presence of this gene in surface waters is indicative of the aerobic heterotrophic metabolism as part of the TCA cycle. In the March and December predictions, malate dehydrogenase shows a considerable reduc-tion in relative abundance within the same spatial region as CO dehydrogenase (Figure 2), which may be indicative of a region of reduced oxygen potential leading to the selection of the reductive acetyl Co-A

carbonyl pathway over the TCA cycle. The Orders Rhodobacteriales and Flavobacteriales show the strongest correlation with the changing relative abundance of malate dehydrogenase (0.425 and 0.409, respectively).

The predicted relative abundances of these three genes, along with 27 others (Supplementary Figure S2) were used to generate a prediction of CO_2 turnover (PRMT score) for each of the 5904 grid cells for the 4 weekly averages (Figure 2). Over the whole WEC, predicted CO_2 accumulation (production; negative PRMT score) was highest in March, which corresponds to an increase in heterotrophic activity in the channel as temperatures increase, phytoplankton bloom and bacteria take advantage of inorganic and organic nutrient-rich waters (Southward *et al.,* 2005; Smyth *et al.,* 2009). As these are weekly averages, it is likely that this prediction reflects a post-phytoplankton bloom condition. Meanwhile, also during March the waters in the southern Irish Sea and off the coast of Cherbourg, France, are predicted to have increased CO_2 consumption, possibly resulting from increased cyanobacterial photosynthetic activity during isolated spring blooms, which presents another example of an opportunity to predict events that could direct empirical evidence gathering. During June, the whole channel starts to show increased CO_2 consumption, which is potentially due to an increase in primary productivity. The greatest consumption (positive PRMT scores) was predicted off the south coast of England in June, which have also been shown to be highly productive (Smyth *et al.,* 2009). During September, this CO_2 consumption maximum spreads through the central channel, and then by December, the central Channel is starting to return to CO_2 production, and therefore a potential dominance of heterotrophy, while the southern Irish Sea and North Coast of France still maintain relatively positive PRMT scores, suggesting CO_2 consumption and the domination of photosynthetic activity. It must be noted that as PRMT scores are relative, all calculations result from a balance between mostly bacterial autotrophic and heterotrophic processes. During March, the accumulation of CO_2 is well distributed, suggesting a dominance of heterotrophic activity. However, there are isolated locations in the southern part of the Irish Sea and in the eastern region of the WEC, which have slightly positive PRMT scores for CO_2 that suggest some catabolism, and hence could be indicative of small phytoplankton blooms. The PRMT scores for CO_2 do not mirror the similarity in the relative abundances of genes coding cyanase, and CO and malate dehydrogenase observed during March and December suggests that the PRMT algorithm, which for CO_2 is based on the changing relative abundance of genes coding 30 different enzymes (Supplementary Figure 2), is capturing a complex emergent property of the differential abundances of this combination of genes.

Extrapolating system-scale predictions across multiple years

Available SIM data (weekly averages from January 2007 to April 2010; Supplementary Figure 1) were used to generate predictions of the changing PRMT scores for 2494 metabolites over 172 weeks, which were clustered by K-means to identify six highly significant groups of inter-annual PRMT score variance (Supplementary Figure 3). These represent calculations of weekly averages, as a channel-wide average with standard deviations for all 5904 grid cells. The six clusters were defined by different patterns in predicted turnover of the constituent metabolites (Supplementary Figure 3i-vi); (i) 104 metabolites that showed accumulation in winter and consumption in summer; (ii) 114 metabolites that showed consumption in winter and accumulation in summer; (iii) 103 metabolites that showed a long period of consumption over summer and a short period of accumulation in the winter; (iv) 163 metabolites that showed a long period of accumulation over summer and a short period of consumption in the winter; (v) 1189 metabolites that showed much smaller PRMT scores than clusters i–iv that generally showed slightly more consumption in the winter than summer; and finally (vi) 817 metabolites that similarly showed smaller PRMT scores than clusters i–iv, and that generally showed slightly more consumption in the summer than the winter. To highlight the potential biological importance of each cluster, we sub-selected a representative metabolite for each (Figure 3). Importantly, with over 2494 metabolites to choose from, this selection is somewhat arbitrary, and influenced by the interests and experience of the authors; however, we selected three metabolites, which show novel patterns that may require experimental and observational validation but highlight the potential of these models to generate novel findings and hypotheses; and three metabolites that validate existing knowledge about their response to seasonal stimuli.

Trimethylamine *N*-oxide (TMAO; Figure 4a), which is a common eukaryotic osmolyte, represented cluster (i) (Supplementary Figure 3). Marine bacteria reduce TMAO to trimethylamine (TMA), which is the cause of the fishy smell in food spoilage and body odor. TMAO reduction and TMA oxidation have an important role in anaerobic bacterial respiration (Barrett and Kwan, 1985). Our model suggests that TMAO is significantly reduced in the summer months (consumption), and that TMA is oxidized to TMAO resulting in accumulation during the winter months, with a sudden increase in catabolism of TMAO each spring, potentially due to the rapid increase in zooplankton biomass in the channel (Smyth *et al.,* 2009). In methylotrophic bacteria, TMA is oxidized to TMAO as a carbon, nitrogen and energy source, TMAO is then demethylated to dimethylamine and formaldehyde, and dimethylamine is ultimately oxidized to CO_2 (Barrett and Kwan, 1985). To test these predictions,

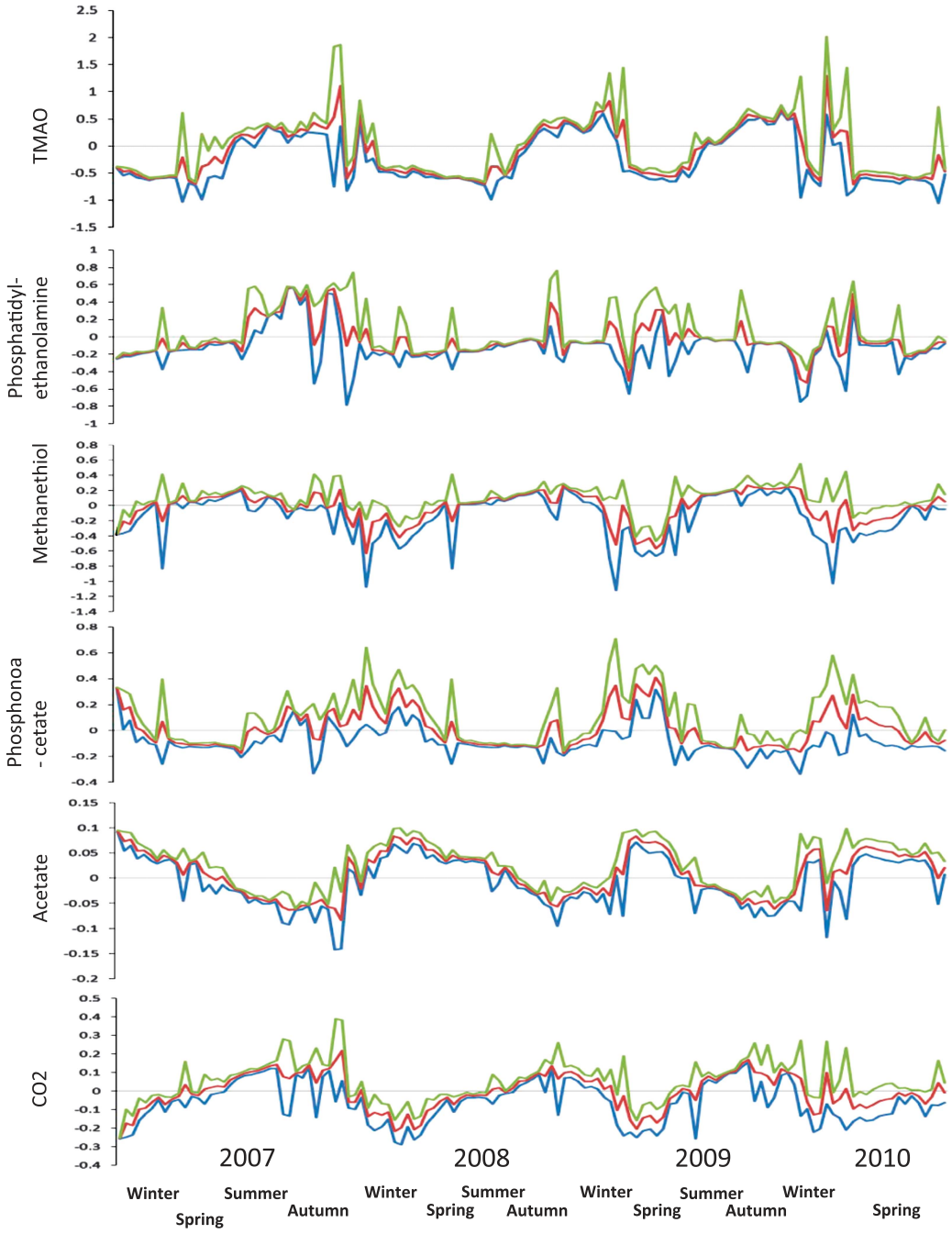

Figure 3 Temporal extrapolation of the relative abundance of microbial orders, and predicted relative metabolic turnover scores for six key marine metabolites. Data are presented as weekly average of 5904 (49 km^2) grid cells across the whole English Channel. Standard deviation is calculated from seven days' of whole English Channel predictions.

we would need to explore the abundance of genes and enzyme activities relating to TMAO metabolism across a trophic gradient (for example, coastal to oligotrophic open ocean). To our knowledge, this is the first report detailing the seasonal and inter-annual variability in TMAO metabolism, which suggests that TMAO may be an important carbon, nitrogen and energy source for microbial metabolism during the summer in the channel. There are periods of high variability in PRMT scores for TMAO (for example, spring and fall 2007, summer 2008, spring 2009, spring 2010); we chose to explore the spatial variance for December 2009, which showed an increase in consumption in isolated regions of the southwest coast of England, and the North coast of France, whereas the rest of the channel had little variation (Figure 4a). Although some periods of apparent instability in the PRMT scores can be explained by the temporal anomalies in the SIM data, which occasionally result in a

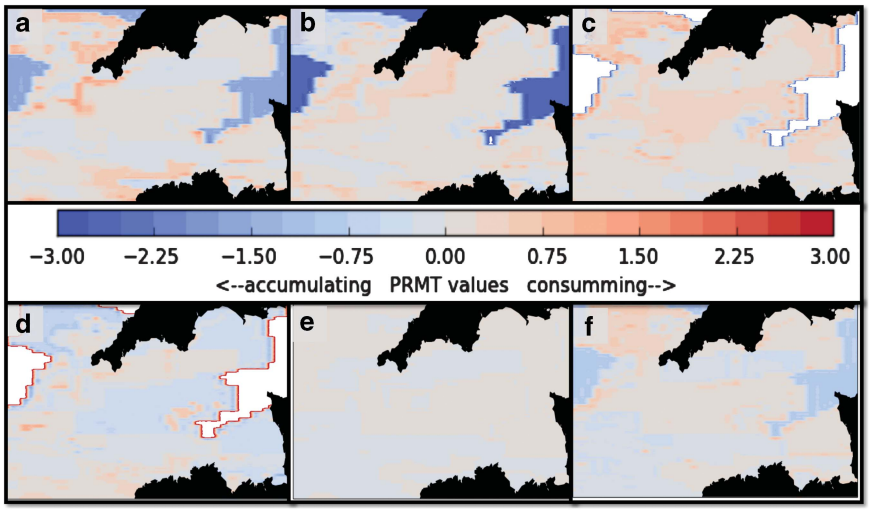

Figure 4 Spatial variation in the predicted turnover of representative metabolites for the six clusters based on a monthly average for December 2009, when each has a considerable discord predicted scores across the channel. (**a**) Trimethylamine N-oxide (TMAO); (**b**) phosphatidylethanolamine; (**c**) methanethiol; (**d**) phosphonoacetate; (**e**) acetate; (**f**) carbon dioxide (CO_2). Regions of deep blue or white repeated for TMAO, phosphatidylethanolamine, methanethiol, phosphonoacetate and CO_2 represent regions where the PRMT scores vary at a very fine scale. To visualize the majority of the channel, these regions of observation are sacrificed.

longer than usual gap between data points (Supplementary Figure 1), none of the high-variability events noted for TMAO fall within these periods. We therefore hypothesize that localized phytoplankton and zooplankton blooms within the channel caused significant regionalization of TMAO metabolism, and generate pockets of bacteria that exploit this resource. This could lead to localized blooms of bacteria within anaerobic micro-niches associated with fecal matter or decaying organic matter associated with a collapsing phyto-/zooplankton bloom (Ditchfield *et al.,* 2012).

Phosphatidylethanolamine (PE; Figure 3b) represents cluster (ii). PE is often associated with heterotrophic bacteria in marine surface waters (Popendorf *et al.,* 2011), and our model predicts that it is generally consumed during the winter and accumulated during the summer. We hypothesize that this is due to a dramatic increase in bacterial biomass during the summer months with more than 1 million cells per ml (Gilbert *et al.,* 2012). Our model predicts many periods of highly variable PRMT scores for PE (Figure 3), some of which correlate with the temporal anomalies in SIM data (Supplementary Figure 1). However, periods through fall-spring 2007, winter-spring 2009 and winter-spring 2010 show extensive variability; again we focused on December 2009 (Figure 4b), which showed an inverse of the spatial turnover scores shown by TMAO, which is unsurprising because these fall into two contrasting clusters (Supplementary Figure 3). One way to explain the inverse relationship between TMAO and PE is that when localized blooms of phyto- and zoo-plankton decay, a bloom of TMAO-reducing bacteria can occur in the anaerobic microniches leading to an increased consumption of TMAO, but also an increased accumulation of PE due to the increase in bacterial biomass. However, this hypothesis remains to be tested.

In cluster (iii), methanethiol (Figure 3c) represents a key intermediary product in the degradation of algal osmolyte dimethylsulfoniopropionate (DMSP). The model predicts general consumption of methanethiol during the spring, summer and fall, and then a short phase of accumulation of this metabolite during December and January each year, which corresponds nearly precisely with measured concentrations of total DMSP in the WEC (Archer *et al.,* 2009). These predictions represent the first time it has been possible to track this important intermediate in marine surface sulfur metabolism over multiple years. In December 2009, high spatial variance corresponds to increased consumption in the Southern Irish Sea, and in the central channel (Figure 4c), whereas the remainder of the channel shows isolated regions of accumulation, which we hypothesize is due to regionalized blooms of DMSP-degrading bacteria.

In cluster (iv), phosphonoacetate (Figure 3d) represents organic phosphate metabolism, with bacterial degradation of the recalcitrant carbon–phosphorus bond as a source of acetate and phosphate (Gilbert *et al.,* 2009); the model confirms previous findings that microbial catabolism of phosphonoacetate occurs primarily during the winter months, which suggests that bacteria are exploiting a metabolic niche during these months when inorganic phosphate is in high concentrations (Smyth *et al.,* 2009). However, December 2009 shows extensive variability in PRMT scores, which may possibly be explained by the extensive regionalization of phosphonoacetate catabolism across the channel, which could result from isolated plankton blooms (Figure 4d).

Cluster (v) is represented by acetate (Figure 3e), a microbial fermentation product oxidized by aerobic heterotrophic bacteria (as an electron donor) to CO_2 via the citric acid cycle, or by anaerobes via the anaerobic carbonyl branch of the reductive acetyl-CoA pathway (Thauer *et al.,* 1989), in which CO-dehydrogenase has a key role (Hügler and Sievert, 2011). As a key metabolite involved in aerobic energy production, which is both rapidly consumed and generated, the low PRMT scores are not surprising, as genes for both consumption and production of acetate are likely to be found in similar relative abundances in this environment, which would lead to reduced relative dynamics in the metabolic turnover across seasons. However, the relative increase in consumption of acetate over winter results from a potential relative decrease in fermentative-generation of acetate during this period, which may then become the dominant process in summer, because of an extensive pulse of organic matter from phytoplankton generation. This hypothesis is supported by existing understanding of acetate metabolism in marine surface waters (Thauer *et al.,* 1989). Spatial variance in PRMT-acetate for December 2009 (Figure 4e) suggests that the channel is divided into defined regions with either relatively greater consumption or accumulation (Figure 3e).

In cluster (vi), like cluster (v), the metabolites all have low PRMT scores (Figure 3). Carbon dioxide represents this cluster (Figure 4f), and demonstrates robust inter-annual turnover, with accumulation of CO_2 in the winter, most likely due to a heterotrophic/autotrophic ratio of >0.5, which is then reversed in summer. Spatial variability during December 2009 (Figure 4f) is different from the weekly average in December 2008 (Figure 2). However, the consumption of CO_2 is still high in the southern Irish Sea, but in 2009, the southwest channel shows significantly greater accumulation, suggesting a collapsed phytoplankton bloom and a subsequent respiration bloom. This region of potentially reduced photosynthesis is mirrored in the acetate map (Figure 4e), with greater accumulation of acetate during this time at this location, suggesting a reduction in CO_2 fixation.

Conclusions

Here we present the first microbial distribution model to predict relative metabolic turnover over broad spatial and temporal scales, highlighting the potential of these models to generate hypotheses that direct future sampling to improve the utility of existing data. This advance represents the next generation of bacterial species distribution models, which previously considered only taxonomic diversity (Larsen *et al.,* 2012; Ladau *et al.,* 2013). It is now possible to extrapolate microbial metabolic potential over vast distances and time scales, as long as corresponding environmental data, including data obtained by inexpensive remote-sensing techniques, exist. This provides an alternative to undirected large-scale observational studies, especially because the predictions here generate discrete biogeographic and biogeochemical hypotheses that may then be validated with directed experiments and observational studies.

The model demonstrated the co-occurrence of potentially anaerobic and aerobic metabolism, which was highly localized. It is possible that this could result of sudden regional increases in anaerobic microsites because of decaying biomass after the summer productivity maximum. Another possible explanation is that storm events could lead to the resuspension of sediments and the introduction of anaerobic microorganisms into the surface water column. Although the same modeling approaches can be used for eukaryotes or viruses, we would need complementary data.

Although our model is predicated on remotely sensed satellite imagery linked to *in situ* data, it is also possible to use any environmental raster data sets to drive predictions in different ecosystems. With recent efforts in microbial niche modeling showing promise in extrapolating microbial community structure across the global ocean (Ladau *et al.,* 2013) and continental soils (Fierer *et al.,* 2013), it is possible that such models could also help to extrapolate metabolic potential across similar scales. As such, this modeling represents a unique opportunity to predict microbially mediated biogeochemical processes, and even, in the future, infer how they will respond to changes in climate.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This work was supported by the US Department of Energy under Contract DE-AC02-06CH11357 and by the Howard Hughes Medical Institute. We also thank the anonymous reviewers from multiple journals who significantly helped us to revise and refine this work to improve the clarity and impact.

References

Archer S, Cummings D, Llewellyn C, Fishwick J. (2009). Phytoplankton taxa, irradiance and nutrient availability determine the seasonal cycle of DMSP in temperate shelf seas. *Mar Ecol Prog Ser* **394**: 111–124.

Barrett EL, Kwan HS. (1985). Bacterial reduction of trimethylamine oxide. *Annu Rev Microbiol* **39**: 131–149.

Brewin RJW, Lavender SJ, Hardman-Mountford NJ, Hirata T. (2010). A spectral response approach for detecting

dominant phytoplankton size class from satellite remote sensing. *Acta Oceanol Sin* **29**: 14–32.
Caporaso JG, Paszkiewicz K, Field D, Knight R, Gilbert JA. (2012). The Western English Channel contains a persistent microbial seed bank. *ISME J* **6**: 1089–1093.
Carder KL, Chen FR, Cannizzaro JP, Campbell JW, Mitchell BG. (2004). Performance of the MODIS semi-analytical ocean color algorithm for chlorophyll-a. *Adv Space Res* **33**: 1152–1159.
Ditchfield A, Wilson S, Hart M, Purdy K, Green D, Hatton A. (2012). Identification of putative methylotrophic and hydrogenotrophic methanogens within sedimenting material and copepod faecal pellets. *Aquat Microb Ecol* **67**: 151–160.
Doney SC, Abbott MR, Cullen JJ, Karl DM, Rothstein L. (2004). From genes to ecosystems: the ocean's new frontier. *Front Ecol Environ* **2**: 457–468.
Fierer N, Ladau J. (2012). Predicting microbial distributions in space and time. *Nat Methods* **9**: 549–551.
Fierer N, Ladau J, Clemente JC, Leff JW, Owens SM, Pollard KS *et al.* (2013). Reconstructing the microbial diversity and function of pre-agricultural tallgrass prairie soils in the United States. *Science* **342**: 621–624.
Follows MJ, Dutkiewicz S. (2011). Modeling diverse communities of marine microbes. *Annu Rev Mar Sci* **3**: 427–451.
Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, Green JL *et al.* (2008). A latitudinal diversity gradient in planktonic marine bacteria. *Proc Natl Acad Sci USA* **105**: 7774–7778.
Gibbons SM, Caporaso JG, Pirrung M, Field D, Knight R, Gilbert JA. (2013). Evidence for a persistent microbial seed bank throughout the global ocean. *Proc Natl Acad Sci USA* **110**: 4651–4655.
Gilbert JA, Field D, Swift P, Newbold L, Oliver A, Smyth T *et al.* (2009). The seasonal structure of microbial communities in the Western English Channel. *Environ Microbiol* **11**: 3132–3139.
Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B *et al.* (2010). The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. *PLoS One* **5**: e15545.
Gilbert JA, Steele JA, Caporaso JG, Steinbrück L, Reeder J, Temperton B *et al.* (2012). Defining seasonal marine microbial community dynamics. *ISME J* **6**: 298–308.
Gilbert JA, Thomas S, Cooley NA, Kulakova A, Field D, Booth T *et al.* (2009). Potential for phosphonoacetate utilization by marine bacteria in temperate coastal waters. *Environ Microbiol* **11**: 111–125.
Glöckner FO, Joint I. (2010). Marine microbial genomics in Europe: current status and perspectives: Marine microbial genomics in Europe. *Microb Biotechnol* **3**: 523–530.
Graetz RD. (1990). Remote sensing of terrestrial ecosystem structure: an ecologist's pragmatic view. In: Hobbs RJ, Mooney HA (eds) *Remote Sensing of Biosphere Functioning* Vol. 79. Springer New York: New York, NY, pp. 5–30 http://www.springerlink.com/index/10.1007/978-1-4612-3302-2_2. Accessed on 25 November 2013.
Hügler M, Sievert SM. (2011). Beyond the Calvin cycle: autotrophic carbon fixation in the ocean. *Annu Rev Mar Sci* **3**: 261–289.
Jørgensen BB, Boetius A. (2007). Feast and famine–microbial life in the deep-sea bed. *Nat Rev Microbiol* **5**: 770–781.
Kamennaya NA, Post AF. (2011). Characterization of cyanate metabolism in marine Synechococcus and Prochlorococcus spp. *Appl Environ Microbiol* **77**: 291–301.
Kamennaya NA, Post AF. (2013). Distribution and expression of the cyanate acquisition potential among cyanobacterial populations in oligotrophic marine waters. *Limnol Oceanogr* **58**: 1959–1971.
Kilpatrick KA, Podestá GP, Evans R. (2001). Overview of the NOAA/NASA advanced very high resolution radiometer Pathfinder algorithm for sea surface temperature and associated matchup database. *J Geophys Res* **106**: 9179.
Ladau J, Sharpton TJ, Finucane MM, Jospin G, Kembel SW, O'Dwyer J *et al.* (2013). Global marine bacterial diversity peaks at high latitudes in winter. *ISMEJ* **7**: 1669–1677.
Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA *et al.* (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* **31**: 814–821.
Larsen PE, Collart FR, Field D, Meyer F, Keegan KP, Henry CS *et al.* (2011). Predicted relative metabolomic turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset. *Microb Inform Exp* **1**: 4.
Larsen PE, Field D, Gilbert JA. (2012). Predicting bacterial community assemblages using an artificial neural network approach. *Nat Methods* **9**: 621–625.
Martin-Cuadrado A-B, Ghai R, Gonzaga A, Rodriguez-Valera F. (2009). CO dehydrogenase genes found in metagenomic fosmid clones from the Deep Mediterranean Sea. *Appl Environ Microbiol* **75**: 7436–7444.
Mason OU, Scott NM, Gonzalez A, Robbins-Pianka A, Bælum J, Kimbrel J *et al.* (2014). Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. *ISMEJ* **8**: 1464–1475.
Paul Bissett W, Schofied O, Mobley CD, Crowley MF, Moline MA. (2001). Optical remote sensing techniques in biological oceanography. In Paul J (ed.), *Methods in Marine Microbiology,* Vol 30. Academic Press: London, pp 519–538.
Pfeil B, Olsen A, Bakker DCE, Hankin S, Koyuk H, Kozyr A *et al.* (2013). A uniform, quality controlled Surface Ocean CO_2 Atlas (SOCAT). *Earth Syst Sci Data* **5**: 125–143.
Popendorf KJ, Lomas MW, Van Mooy BAS. (2011). Microbial sources of intact polar diacylglycerolipids in the Western North Atlantic Ocean. *Org Geochem* **42**: 803–811.
Schmidt M, Lipson H. (2009). Distilling free-form natural laws from experimental data. *Science* **324**: 81–85.
Scott NM, Hess M, Bouskill NJ, Mason OU, Jansson JK, Gilbert JA. (2014). The microbial nitrogen cycling potential is impacted by polyaromatic hydrocarbon pollution of marine sediments. *Front Microbiol* **5:** http://www.frontiersin.org/Aquatic_Microbiology/10.3389/fmicb.2014.00108/abstract. Accessed on 24 April 2014.
Smith VA, Yu J, Smulders T, Hartemink AJ, Jarvis ED. (2005). Computational Inference of Neural Information Flow Networks. *PLoS Comput Biol* **2**: e161.
Smyth TJ, Fishwick JR, AL-Moosawi L, Cummings DG, Harris C, Kitidis V *et al.* (2009). A broad spatio-temporal view of the Western English Channel observatory. *J Plankton Res* **32**: 585–601.
Southward AJ, Langmead O, Hardman-Mountford NJ, Aiken J, Boalch GT, Dando PR *et al.* (2005). Long-term

oceanographic and ecological research in the Western English Channel. *Adv Mar Biol* **47**: 1–105.

Thauer RK, Möller-Zinkhan D, Spormann AM. (1989). Biochemistry of acetate catabolism in anaerobic chemotrophic bacteria. *Annu Rev Microbiol* **43**: 43–67.

Toseland A, Daines SJ, Clark JR, Kirkham A, Strauss J, Uhlig C *et al.* (2013). The impact of temperature on marine phytoplankton resource allocation and metabolism. *Nat Clim Change* **3**: 979–984.

Wang X, Mannaerts CM, Yang S, Gao Y, Zheng D. (2010). Evaluation of soil nitrogen emissions from riparian zones coupling simple process-oriented models with remote sensing data. *Sci Total Environ* **408**: 3310–3318.

Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* **20**: 3594–3603.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)