Molecular Systems Biology 7; Article number 473; doi:10.1038/msb.2011.6
Citation: ***Molecular Systems Biology*** **7:473**
© 2011 EMBO and Macmillan Publishers Limited All rights reserved 1744-4292/11
www.molecularsystemsbiology.com

Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data

Jeroen Raes[1,2], Ivica Letunic[1], Takuji Yamada[1], Lars Juhl Jensen[1,3] and Peer Bork[1,4,*]

[1] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany, [2] Molecular and Cellular Interactions Department, VIB – Vrije Universiteit Brussel, Brussels, Belgium, [3] NNF Center for Protein Research, Copenhagen, Denmark and [4] Max Delbrück Center for Molecular Medicine, Berlin-Buch, Germany

* Corresponding author. Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, Heidelberg 69117, Germany. Tel.: +49 6 221 387 8526; Fax: +49 6 221 387 8517; E-mail: bork@embl.de

Received 4.5.10; accepted 25.1.11

Using metagenomic 'parts lists' to infer global patterns on microbial ecology remains a significant challenge. To deduce important ecological indicators such as environmental adaptation, molecular trait dispersal, diversity variation and primary production from the gene pool of an ecosystem, we integrated 25 ocean metagenomes with geographical, meteorological and geophysicochemical data. We find that climatic factors (temperature, sunlight) are the major determinants of the biomolecular repertoire of each sample and the main limiting factor on functional trait dispersal (absence of biogeographic provincialism). Molecular functional richness and diversity show a distinct latitudinal gradient peaking at 20°N and correlate with primary production. The latter can also be predicted from the molecular functional composition of an environmental sample. Together, our results show that the functional community composition derived from metagenomes is an important quantitative readout for molecular trait-based biogeography and ecology.

Molecular Systems Biology **7:** 473; published online 15 March 2011; doi:10.1038/msb.2011.6
Subject Categories: bioinformatics; functional genomics
Keywords: ecosystems biology; environmental genomics; metagenomics; microbiology; molecular trait-based ecology

This is an open-access article distributed under the terms of the Creative Commons Attribution Noncommercial Share Alike 3.0 Unported License, which allows readers to alter, transform, or build upon the article and then distribute the resulting work under the same or similar license to this one. The work must be attributed back to the original author and commercial use is not permitted without specific permission.

Introduction

Microbial communities have a central role in global environmental processes and the Earth's biogeochemistry by cycling nutrients and fixing carbon (Falkowski *et al*, 1998). However, because of the complexity of these communities and the lack of culturability of most of its members, the molecular and ecological details as well as influencing factors of these processes are still poorly understood. Environmental shotgun sequencing (metagenomics) has the potential to start unraveling the underlying complex interspecies ecological interactions and metabolic networks, by quantification of the molecular functions ('parts lists') of all microbial communities on Earth (Tringe and Rubin, 2005a; Raes and Bork, 2008). However, despite a wide range of published metagenomics studies (see Liolios *et al*, 2006; Raes *et al*, 2007; Wooley *et al*, 2010, for an overview), our knowledge of the variation, functioning and ecology of complex microbial ecosystems remains limited, mostly because the resulting 'parts lists' could not be put into sufficiently detailed environmental context. Although previous studies have shown that the environment has an influence on the parts list of various communities, the extent of this effect and the relative importance of a broad range of different environmental factors (climate, nutrients, physicochemical parameters and so on) is unknown (Tringe *et al*, 2005b; DeLong *et al*, 2006; Dinsdale *et al*, 2008; Kunin *et al*, 2008; Gianoulis *et al*, 2009) or was investigated with a focus on single species (Johnson *et al*, 2006) or specific gene families (Patel *et al*, 2010). This said, recent models predicting nutritional strategy from metagenomic data show great promise toward the understanding of some of these relationships (Lauro *et al*, 2009). Also, as microbial biogeography and ecology studies have mostly focused on phylogenetic patterns, little is known about the role of molecular traits (i.e., the genes and their products) in these matters (Martiny *et al*, 2006; McGill *et al*, 2006; Green *et al*, 2008). Likewise, the role of molecular trait variation in important ecosystemic processes such as global primary production is far from clear (Falkowski *et al*, 1998). To start addressing these issues, we investigated the feasibility of molecular trait-based ecology by integrating large-scale marine metagenomics data with geochemical, meteorological and ecological measurements and used this information to investigate (i) the relationship between environment and functional community composition (the metagenome-derived gene/pathway repertoire of an ecosystem), (ii) the factors influencing functional dispersal (defined here as the functional effects of species dispersal as well as horizontal gene transfer- and phage-mediated gene flow), i.e., the movement of functional traits through geographical space, (iii) the interplay between functional composition and primary production and (iv) the geographic variation in global functional diversity and its consequences. The various correlations we found, despite various imaginable limitations of environmental sequence data (see further), thereby indicate that molecular functional composition, as derived from metagenomes, can serve as a powerful marker and predictor of ecological processes.

Results and discussion

We utilized the Global Ocean Survey (GOS) which is, at the time of writing, the largest published metagenomic study of a single environment (excluding host-associated communities; Qin *et al*, 2010), gathering ocean surface samples from a transect around the globe (Rusch *et al*, 2007). Although it has some drawbacks (e.g., size fractionation excluding eukaryotic plankton; only dominant species are sampled), it still constitutes a unique data set to assess the feasibility of molecular trait-based ecological studies. We mapped these data onto orthologous groups (OGs) and biochemical pathways (allowing multi-level functional interpretations and to overcome undersampling issues in gene-based analyses; Harrington *et al*, 2007), and linked pathways to species, when possible, in order to interpret correlations in the context of the phylogenetic composition of the community (see Materials and methods). Then, we complemented the metadata gathered in the GOS project (which we studied in Gianoulis *et al* (2009)) by projecting a broad range of geophysicochemical (e.g., nitrate, phosphate, oxygen measurements, ocean mixing), geographical (latitude, longitude, depth) and meteorological data (e.g., temperature, sunlight) as well as ecological information (primary production) from publically available resources (see Materials and methods and references therein) onto 25 published metagenome sampling points (Rusch *et al*, 2007) using the sampling time and coordinates (see Materials and methods; Supplementary Table S1 and Supplementary Figure S1).

Functional community composition is mainly driven by climatic factors

Previous studies have shown a clear impact of the environmental conditions on the functional composition of microbial communities (Tringe *et al*, 2005b; DeLong *et al*, 2006; Dinsdale *et al*, 2008; Kunin *et al*, 2008; Gianoulis *et al*, 2009). To investigate which environmental conditions are the main drivers in this process, we applied previously established techniques (Gianoulis *et al*, 2009) to the integrated metagenomic and environmental data. Canonical correlation analysis (CCA; Hotelling, 1936; Gianoulis *et al*, 2009) shows that the overall correlation between environmental factors and various biochemical pathways is high and that numerous pathways have strong correlations with environmental factors (see Figure 1, Supplementary Tables S2-4). Generally, temperature, sunlight, oxygen and CO_2 concentration have the strongest correlations, as they distribute along the dimension with the highest canonical correlation coefficient (CC=0.944, see Figure 1C), while salinity and nutrients contribute more to the second, less significant, dimension (CC=0.875; only one significant module, Supplementary Table S4; see Supplementary Table S5 for more evaluation metrics). This suggests that, for the ocean surface communities analyzed here, climatic factors such as sunlight and temperature (and correlated dissolved oxygen and CO_2 content) are the main determining factors of the functional community composition, whereas nutrient concentrations seem to have less influence, despite their crucial, limiting role in ocean life and productivity (Arrigo, 2005). As phylogenetic composition was reported to be mainly determined by salinity in a survey of various environments (Lozupone and Knight, 2007), we repeated the CCA analysis on the phylogenetic composition of the GOS samples used here (data taken from Biers *et al*, 2009) but found the phylogenetic results to be in agreement with our functional trends (Supplementary Figure S2). Given the range of salinity in our study was much smaller than in the aforementioned study, it could however be that salinity has a role in more extreme concentrations and outside ‘normal’ oceanic samples as studied here. We should also note that these correlations are based on available monthly average values, so that nutrients could still have a larger role in fast adaptations at much shorter timescales (Gilbert *et al*, 2009). Data sets that cover a larger and physicochemically more diverse geographic region can be expected in the near future (e.g., from the Tara Oceans project, http://oceans.taraexpeditions.org) and will provide a higher resolution to refine our initial observations.

Molecular adaptions to environmental conditions

Several correlations between environmental factors and abundance of metabolic pathways as revealed by CCA were further confirmed using pairwise correlation analysis and linear regression (see Materials and methods). Some of these confirm common knowledge of ocean microbial processes. For instance, a strong positive correlation was found for both temperature and hours of sunlight with the abundance of genes encoding the photosynthetic machinery (Figure 1A). Not only the light capturing complexes (PSI, PSII, CytB6) but also modules involved in oxidative phosphorylation and CO_2 fixation were positively correlated with temperature, confirming the greater dependence on (photo-) autotrophic processes in sunny, warm and nutrient-poor tropical areas (mainly driven by samples from the Gulf of Mexico and, to a lesser extent, the tropical Pacific). The phylogenetic CCA analysis confirms this observation, by showing strong association of photosynthetic groups such as cyanobacteria with temperature (Supplementary Figure S2).

© 2011 EMBO and Macmillan Publishers Limited

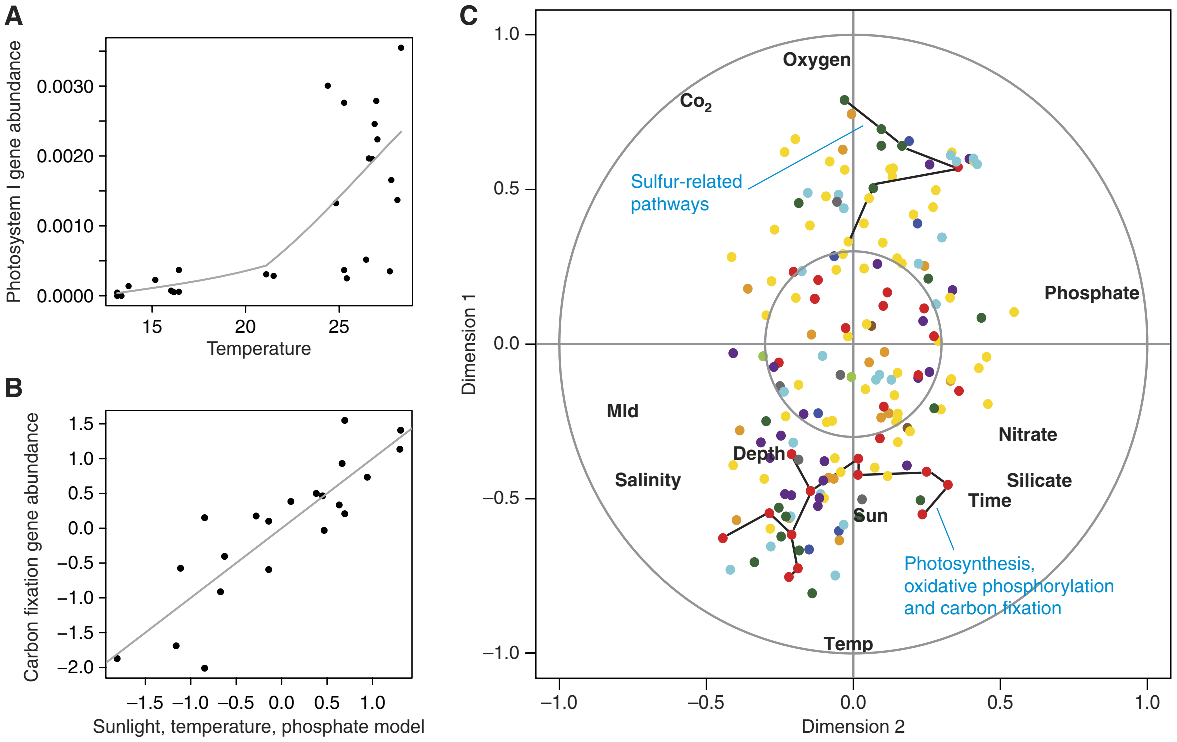


Figure 1 Correlations between metabolic pathway abundances and environmental conditions deduced from the ocean samples in this study, at various levels of model complexity (see Materials and methods): (**A**) 'One-to-one' pairwise correlation (*P*=0.001) between the abundance of photosystem I genes with average monthly water temperature. (**B**) 'One-to-many' linear model of average monthly water temperature, phosphate concentration and hours of sunlight correlating with carbon fixation gene abundance (R^2=0.70). (**C**) 'Many-to-many' regularized canonical correlation analysis ordination plot showing the correlation between all environmental variables (text labels; see Materials and methods) and pathway modules (colored dots). The distance between two variables on the plot and their distance from the center point indicates the strength of their correlation and their contribution to explaining the global correlation (i.e., their structural correlation in each dimension given on the respective axes: first dimension, vertical, second, horizontal; see Materials and methods). The overall canonical correlation is high (canonical correlation=0.944 in the first dimension), and the two first dimensions explain 62 and 22% of the total environmental and metagenomic variation, respectively, emphasizing the strong correlation between the climatic factors and functional community composition on the first dimension. Module colors indicate their broad functional classes: yellow, amino acid metabolism; orange, central metabolism; red, energy metabolism, dark green, glycan metabolism; cyan, lipid metabolism; purple, metabolism of other molecules; blue, nucleotide metabolism; brown, replication and repair; light green, transcription; pink, translation; gray, transport system. Highlighted modules are described in more detail in the text. mld, mixed layer depth.

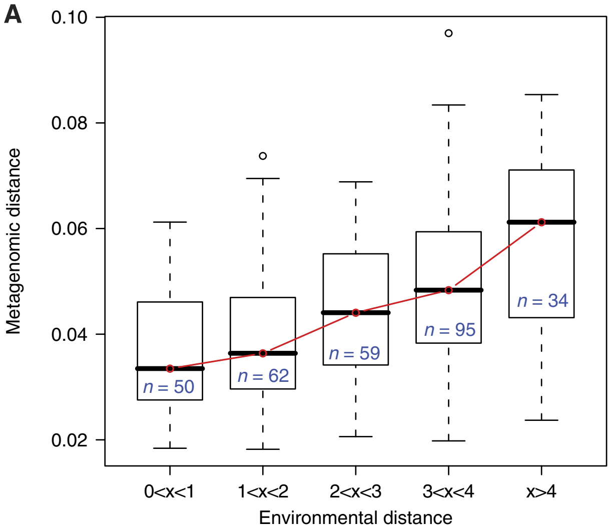
Other examples include a wide range of sulfur-related processes that are negatively correlated to temperature ranging from enzymes linked to sulfate reduction and/or sulfite detoxification (adenylyl sulfate reductases; Meyer and Kuever, 2007) to pathways involved in methionine degradation and breakdown of sulfur-containing glycans. The former seem to originate mainly from SAR11-like species (where they would be reverse-acting to prevent sulfite accumulation during dimethylsulfoniopropionate assimilation (Meyer and Kuever, 2007)), but are also found in various other bacterial groups. The latter include mainly sulfatases from planktomycetes (see Supplementary Figure S3 for phylomapping results), which are suggested to be involved in the initial breakdown of sulfated heteropolysaccharides in heterotrophic carbon recycling (Woebken *et al*, 2007). The observed anticorrelation of these processes with temperature suggests a more general dependence on organic sulfur (despite high ambient sulfate concentrations) in the northern, coastal heterotroph-dominated communities, something which was currently only observed for specific phylogenetic groups (i.e., SAR11; Tripp *et al*, 2008).

Functional biogeography and dispersal are primarily determined by environmental conditions

Having demonstrated that functional composition can be clearly linked to external conditions, we used functional traits derived from metagenomes to study, as a second application, function dispersal and biogeography in ocean samples. Various rRNA-based studies over many years remain undecided on the existence and nature micro-organismal biogeographic patterns (Finlay, 2002; Martiny *et al*, 2006; Telford *et al*, 2006). As evidence accumulates that the set of functional traits, not the rRNA genes, are the true ecological determinants of a microbial species (Gevers *et al*, 2005; McGill *et al*, 2006; Green *et al*, 2008; Hunt *et al*, 2008; Fraser *et al*, 2009), we investigate the use of molecular functional traits as biogeographic markers. Given rampant horizontal gene transfer and, e.g., the frequent observation of bacterial genes in phages, we here use the term 'functional biogeography' to allow for the possibility that the traits themselves disperse irrespective of their original hosts, although it deviates from the strict definition where biogeography is only applicable to the trait-bearing organisms.

© 2011 EMBO and Macmillan Publishers Limited

To determine the distribution of function in geographical space and identify its determining factors, we compared the difference in functional (metagenomic) composition (i) between the samples (as measured by the Bray–Curtis distance metric; see Materials and methods), (ii) to the geographical distance between them and (iii) to the difference in environmental conditions (Martiny *et al*, 2006). As climate is the principle factor influencing the functional composition of the communities studied here, we demonstrate its effect on function dispersal. The functional difference between communities increases with difference in climatological conditions (Figure 2A; Mantel test, $P<0.001$), and this effect remains even if the difference in environmental conditions caused by geographic distance is subtracted (Figure 2B; partial Mantel test, $P=0.01$; see Materials and methods and Supplementary Table S6 for details on tests performed). On the other hand, when we tested for biogeographic provincialism (i.e., a 'distance effect', implying geographic limitations to function dispersal), no significant correlation between physical distance and functional difference could be found when difference in environmental conditions was taken into account (Figure 2B; partial Mantel test, $P=0.1$). This implies that the functional traits available to a specific community have no physical constraints on their dispersal, and their abundances are mainly determined by local, contemporary environmental conditions. In other words, a single functional province seems to exist over hundreds to thousands of kilometers in the surface ocean of the Atlantic and Pacific. These observations provide evidence for a functional equivalent of the (organism-centric) Baas-Becking theory (Baas Becking, 1934), namely 'all *functions* are everywhere, but the environment selects'. One should note that the trend diminishes if not only climatic but also nutritional variables are included (see Materials and methods), suggesting that selection of relevant environmental parameters can lead to the identification of otherwise hidden biogeographic patterns. On the other hand, the same correlation analysis, when performed on the phylogenetic composition (source data from Biers *et al*, 2009), shows a (weak) distance effect (partial Mantel test, $P=0.03$, see Supplementary Table S6 and Materials and methods for details) compatible with previous reports of microbial species-level biogeographic provincialism (Martiny *et al*, 2006; Telford *et al*, 2006). However, the phylogenetic composition seems not to be selected for by environmental factors (partial Mantel test, $P=0.25$, see Supplementary Table S6), suggesting that phylogeny and function are not necessarily coupled. Thus, while for phylogeny more neutral population effects dominate in geographical distinct locations, selection for environmental constraints seems more readily distinguishable in molecular functions.



B

Tested correlation	Mantels′ *r*	*P*-value
Functional (metagenomic) distance versus environmental distance	0.43	<0.001
Functional (metagenomic) distance versus geographical distance	0.38	<0.001
Functional (metagenomic) distance versus environmental similarity, given geographical distance	0.27	0.01
Functional (metagenomic) distance versus geographical distance, given environmental distance	0.12	0.1

Figure 2 The role of environment in the biogeography of functional traits. (**A**) Coupling of metagenomic distance between samples (measured using KEGG metabolic pathway composition; see Materials and methods) with difference in climatological conditions, identifying climate as a primary determinant of function dispersion. (**B**) (Partial) Mantel tests (see Materials and methods) showing that this increase is not due to indirect effects, such as the similarity in environmental conditions between geographically close samples.

Functional community composition can predict primary production

The case studies above exemplify a wealth of apparent metabolic adaptations to environmental conditions and prove the ability to extract those from massive amounts of data. This prompted us, in a third application, to explore the potential of functional composition in predicting other global ecological parameters, such as primary production (the amount of CO_2 fixed by photosynthetic micro-organisms that is available for other trophic levels as organic carbon (Lindeman, 1942). Primary production is a crucial parameter in global carbon cycling and climate change and is calculated from, among others, temperature and sun irradiation measurements (Field *et al*, 1998). We retrieved information on the monthly average primary production for the sampling coordinates (Field *et al*, 1998) and correlated this information to both the functional composition of the microbial community sampled and the environmental conditions of the sampling site. As expected, primary production was predictable from environmental conditions (data not shown). However, also the abundance of several central, core metabolic pathways from the metagenome strongly correlates with primary production (see Supplementary Table S7). Examples include pathway modules involved in nucleotide (dCTP, UMP) and sugar-nucleotide (UDP-glucose/UDP-galactose) biosynthesis (Spearman's $\rho=0.79$, -0.69 and -0.69, respectively), photosynthesis (Photosystem I and II; $\rho=-0.71$ and -0.53, respectively) and oxidative phosphorylation (complex I; $\rho=-0.76$). The negative correlation of photosynthesis and linked energy generation with primary production, at first sight counterintuitive, can be

© 2011 EMBO and Macmillan Publishers Limited

explained by the filter sizes used in the metagenome sampling, which excludes several of the main contributors to primary production (diatoms, dinoflagellates, green algae). In areas of low nutrient concentration and low global primary production, photosynthetic cyanobacteria thrive because of their specific adaptation to these circumstances (Vaulot *et al*, 1995). In areas of high nutrient concentration, where the major eukaryotic determinants of primary production thrive, they could outcompete the photosynthetic bacteria, which are the only ones captured in the GOS data sets, as the cell size fractions of the samples used here did not target protist species. On the other hand, the observed strong correlation between many additional core metabolic pathways of only a particular subfraction of the ecosystem population (bacteria/archaea) and the total measured primary production provides evidence for a robust metabolic coupling between these organisms and the carbon fixing eukaryotes in the ocean surface. This hypothesis is reinforced by the fact that linear models consisting of only abundances of the most correlated metabolic Kyoto Encyclopedia of Genes and Genomes (KEGG) maps (Kanehisa *et al*, 2008) were almost equally successful at predicting primary production as the external environmental measures used (R^2_{env}=0.91 versus R^2_{met}=0.87; see Supplementary Figure S4 and table S8). This observation confirms a tight coupling of functional composition of the sampled bacterial community and global primary production at that location. Furthermore, this observation suggests that functional properties, identifiable through metagenomic studies can be used to predict various community properties and processes that are hard to measure in environmental settings.

Functional richness and diversity show a distinct latitudinal gradient and are linked to primary production

In addition to the functional and phylogenetic composition of microbial communities, also global estimators of the richness (e.g., number of species) and diversity (richness, corrected for population structure) of an ecosystem are widely used to investigate and understand ecosystem properties (Colinvaux, 1973). For instance, community productivity has been repeatedly correlated to species richness. However, depending on the environmental circumstances, also negative correlations have been observed (Kondoh, 2001; Loreau *et al*, 2001). In microbial ocean communities, no correlation was found so far (Fuhrman *et al*, 2008). At the macroscopic functional level, positive correlations between richness in macroscopic traits (e.g., plant functional groups in a field study) and productivity have been reported (Tilman *et al*, 1997; Loreau *et al*, 2001). We therefore investigated whether estimators of trait diversity at the molecular level can provide relevant alternatives to study microbial communities, as they consider the direct functional units (genes) that determine the properties inherent to the ecosystem. One comprehensive measure of molecular functional richness is the extrapolated richness in OGs present in a metagenomic sampling of a community, which quantifies the ‘breadth’ of the functional potential of the ecosystem at hand (Raes and Bork, 2008). When also the evenness of the functional distribution can be taken into account, functional diversity can be determined (Raes and Bork, 2008). Applied to the data used here, we observe a clear peak in functional richness at 20 degrees north along the north–south transect of the GOS expedition (for details see Figure 3A and B). Functional diversity shows a similar but less pronounced pattern (see Supplementary Figure S5), and similar results are obtained for gene family richness (see Supplementary Figure S6). These results are in line with previous reports of latitudinal species richness gradients, although in these studies the gradient has a more noisy distribution (Fuhrman *et al*, 2008) or the position of the peak cannot be observed because of the relatively low number of samples (Pommier *et al*, 2007).

We also observe a significant negative correlation between functional diversity and primary production (see Figure 3C; also observed for functional richness, data not shown). These results seem in disagreement with plant macroscopic functional trait richness, which positively correlates with productivity (Tilman *et al*, 1997; Loreau *et al*, 2001). In addition, published marine bacterial species richness analyses (Fuhrman *et al*, 2008), as well as with our own reanalysis of the phylogenetic composition of these samples (see Supplementary Figure S7), show no or at best a very weak (P=0.28) correlation with primary production. Our results are in line with ecological theory though, namely that in areas of high productivity, a narrow functional niche of organisms (low functional richness) impacts the community and drives the majority of productivity. In low-productivity (nutrition poor) areas, more functional niches are formed because of a higher level of competition for resources, causing a broader global set of functionalities (Krebs, 2001). Given the theoretical support for our results, the weakness of the phylogenetic signal when derived from PCR bias-free metagenomic sequencing (von Mering *et al*, 2007a), and in line with the increasing realization that the ecological function of a microbial organism is a true indicator of a species (Gevers *et al*, 2005; Hunt *et al*, 2008; Fraser *et al*, 2009), we conclude that functional diversity estimators (such as richness) and composition might be more relevant and immediate indicators of the link between diversity and ecosystem processes than species-based measures.

Metagenome-derived functional composition is an important tool for molecular ecology and biogeography research

Taken together, this study extends previous work (Gianoulis *et al*, 2009) through a comprehensive integration of metagenomic data with a broad range of quantitative environmental factors extrapolated from a variety of measures independent from the sampling, allowing the identification of climatic factors as the drivers behind functional community composition and functional biogeography. It is further the first attempt to establish the molecular functional repertoire of a metagenomic sample as indicator and predictor of ecological parameters. The metagenomes used here (Rusch *et al*, 2007) are, of course, only snapshots of the functional potential of the environment and probably still give a biased view on the total biodiversity (e.g., by the sampling method, sequence

© 2011 EMBO and Macmillan Publishers Limited

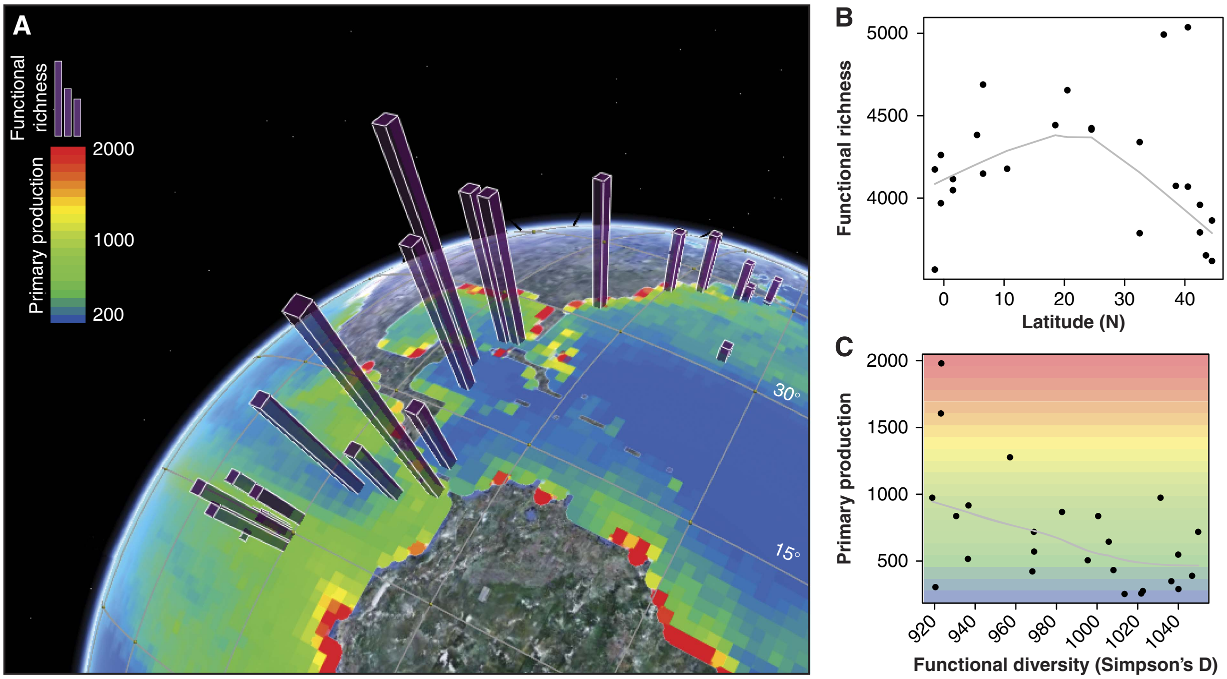


Figure 3 Variation of functional richness and diversity, and its coupling to primary production. (**A**) Global view of primary production for the sampled region (ocean coloring) and functional richness for GOS sampling locations used in this study (3D bars), showing the peak in richness at ±20 degrees North (two outliers were removed for visualization purposes, but are present in (**B**), showing the full, quantitative data plotted against latitude), (**C**) functional diversity negatively correlates with primary production (Spearman's $\rho = -0.49$; $P = 0.01$). Trend lines are Lowess fitted lines with smoothing parameter $f = 0.7$.

coverage, filtering and so on). Future dedicated experiments and time series data will greatly improve causal inferences beyond the simple logic used here (environment influences community composition and functioning, which influences ecosystem processes such as primary production). The environmental and ecological variables used are mostly monthly averages of the region, and not exact measurements at the time of sampling. Furthermore, multiple measures had to be taken to avoid parameter overfitting and other testing artifacts (see Materials and methods). However, despite these drawbacks and complications, we see clear, significant trends relating the environment and ecological parameters to metagenomic gene abundances (to the extent of having predictive power), providing first insights into the coupling between the functional traits available to the community, the environmental context and the ecosystem processes that occur. We see this as proof-of-principle that molecular functional composition can be used in various other environmental settings such as the human microbiome, where it could be integrated with clinical data to study the molecular ecology and temporo-spatial variation of the 'human' ecosystem.

Materials and methods

GOS data collection and sequence preprocessing

For this study, we filtered the data from the first phase of the GOS expedition to keep only those sites that used a 0.1–0.8 μm filter size (i.e., majority prokaryote samples). In addition, atypical samples from mangroves, estuaries, salt lakes, large depth and so on were not included in the analysis, nor was the sample of Sargasso Sea station 11, because it is suspected of contamination (Mahenthiralingam *et al*, 2006), and that of Cabo Marshall because of a considerable number of missing environmental data points (see below). For the remaining 25 sites (Supplementary Table S1), protein sequence data were downloaded from CAMERA (Seshadri *et al*, 2007). Peptides were mapped to sites based on the read-to-scaffold and ORF-to-scaffold mappings available at the same database, using previously established methods (Tringe *et al*, 2005b; Kunin *et al*, 2008; Gianoulis *et al*, 2009).

Geographical, environmental and ecological metadata mapping

Sampling longitude, latitude, depth, date and time were extracted from the CAMERA database (Seshadri *et al*, 2007). Average monthly dissolved oxygen, phosphate, nitrate, silicate, temperature and salinity were extracted from the 1 degree gridded data available at the World Ocean Atlas (WOA05) resource (Boyer *et al*, 2006) for the geographically closest available data points for the selected samples, using the longitude, latitude, month and year of sampling. Gridded dissolved carbon dioxide data from the GLODAP resource (Key *et al*, 2004), and mixed layer depth data (estimated based upon a 0.5 °C surface water temperature change (mld_pt); Monterey and Levitus, 1997) were extracted using the Ocean Data Viewer software (http://odv.awi.de/en/home/). Average monthly primary production data, as estimated by the Vertically Generalized Production Model (VGPM) (Behrenfeld *et al*, 2006), was downloaded from the ocean productivity resource (http://www.science.oregonstate.edu/ocean.productivity/index.php). Monthly average sunshine fraction (the percentage of time when bright sunshine is recorded during the day) data was obtained from the local climate estimator (LocClim) resource (Grieser, 2002), based upon default interpolation settings. Environmental parameters for which both on-site (Rusch *et al*, 2007) as well as monthly averages were available (temperature, salinity) were compared and showed strong consistency among samples (Supplementary Figure S1).

Functional annotation and pathway assignment

The 111 KEGG maps, 141 modules and 191 operons were assigned as in Tringe *et al*, 2005b; Kunin *et al*, 2008; Gianoulis *et al*, 2009. Module definitions were downloaded from KEGG (Kanehisa *et al*, 2008), and operons were constructed as in von Mering *et al*, 2003 and Kunin *et al*, 2008. For clarity, in the remainder of the text, we use the term pathway to refer to all of these levels. Pathway abundance and presence was measured as in Tringe *et al*, 2005b; Kunin *et al*, 2008; and Gianoulis *et al*, 2009. In brief, predicted protein sequences were searched against the extended database of proteins assigned to OGs and pathways in STRING 7.0 (von Mering *et al*, 2007b), by using BLASTP (Altschul *et al*, 1990), and an OG/pathway was called present when a hit matching 1 of its proteins occurred (with a BLAST score of at least 60 bits; see Supplementary Table S9 for further annotation statistics). Gene frequencies were determined by counting the number of reads contributing to that gene, standardizing for sample size. Likewise, the OG/pathway frequency for each site was assigned by summing the total number of instances of that OG/pathway (i.e., reads mapping to a gene assigned to that pathway) for a particular site and standardizing by total number of assignments for that site to compensate for sample coverage differences. For all correlation analyses, pathways for which the summed count over all sites constituted equal to or less than 0.01% of the total count were removed to avoid artifacts. Correlation was carried out using the relative counts of genes involved in specific metabolic pathways (or modules/OGs). For ease of reading, this is described as a correlation between an environmental factor and the pathway. All results described were also manually scrutinized, and for all case studies, confirmation was sought at multiple levels of resolution (map-module-operon-OG) to reduce artifacts.

Pairwise correlations, linear regression and canonical correlation analysis

Correlation analysis on the extended set of environmental parameters and ecological variables was performed using the same methodology as used in Gianoulis *et al* (2009). In brief, we computed pairwise Spearman correlations between pathway frequencies and environmental variables over all samples (controlling the false discovery rate by Benjamini and Hochberg (1995) correction for multiple testing. Linear models were constructed in two directions: (i) in the case of the prediction of primary production the pathway frequencies acted as dependent variables and environmental conditions the response variables, whereas in (ii) the investigation of the effect of environment on community composition, pathway frequency was treated as the response variable and predicted from environmental factors. For both models, we used stepwise regression analysis (SRA; implementation in the R-stats package) to reduce the number of response variables in the model. To avoid overfitting in (i), we used only the top 15 pathways that showed the highest pairwise correlation (as measured by uncorrected *P*-value) with the environmental feature modeled. Linear models were considered significant at $P<0.05$ for both the total model and the estimate of the variable coefficients. For regressions in both directions, the pathway frequencies were standardized to a mean of 0 and a s.d. of 1. For (i), we used the centered, quantile-normalized primary production data transformed into percentiles to ensure a truly normal distribution and, thus, accurate *P*-values. As the linear model construction procedure did not allow any missing values, we removed all samples with missing environmental data values (21 samples remaining). In (i), a leave-one-out cross-validation procedure was used to assess the behavior of the derived model on samples not used for training. This procedure was used both at the feature selection step as well as the prediction step. First, the set of dependent variables with significant weights in the model were determined for all $(n-1)$ sample combinations using the SRA feature selection approach, and the most frequent set of features was chosen as the final model (see Supplementary Table S8 for results). Next, using these features, parameter estimation was performed on each combination of $(n-1)$ samples and the predicted primary production was compared with the observed value (see Supplementary Figure S4 for results). Regularized CCA was used to identify the set of projections that maximally correlate pathways and environmental variables (Gianoulis *et al*, 2009). Structural CCs (the correlation between the original variable and the canonical variate) were used to estimate the importance of one variable relative to all of the other in the maximization of the correlation between pathways and factors. To test the statistical significance of structural correlations, 1000 randomized data sets (sample labels permuted) were analyzed in the same way as the original data set. The statistical significance was calculated by comparing the observed structural correlations to the distributions obtained from randomization, and a significance threshold of $P<0.05$ after Benjamini and Hochberg (1995) correction for multiple testing was applied.

Functional diversity estimates and biogeography

Functional richness and diversity (Raes and Bork, 2008) were estimated from the OG abundance counts (OG richness) using EstimateS (http://viceroy.eeb.uconn.edu/estimates). Richness was calculated using the Chao1 estimator (Chao, 1984) and for functional diversity, Simpson's diversity index was used (Magurran, 1988). We also investigated the use of other indices (e.g., ACE, Michaelis Menten), but this did not affect results very much (data not shown). In addition, gene family richness was determined from gene family mappings of each sample (data from Yooseph *et al*, 2007).

To investigate the relative importance of environment (climate) and geographic separation, Mantel and partial Mantel tests (based on Pearson's correlations, 10 000 permutations) were performed on environmental similarity, metagenomic similarity and geographic distance matrices, which were calculated using Euclidian distances between the scaled environmental variables, Bray–Curtis distances between pathway frequency matrices and raw haversine distances (in km) between samples, respectively. Calculations were done using the functions in the 'stats', 'ecodist' and 'vegan' libraries in the R-package (www.r-project.org; Goslee and Urban, 2007; Oksanen *et al*, 2008).

Visualization

Visualization of phylogenetic data was performed using iTol (Letunic and Bork, 2007). Primary production and functional richness data were mapped on the globe using google earth (http://earth.google.com).

Species mapping

Proteins from selected pathways were mapped to nodes of the tree of life (Ciccarelli *et al*, 2006) using an in-house perl script, based upon the lowest common ancestor approach (Huson *et al*, 2007). Input data were BLASTp results of the proteins against the STRING 7 database (von Mering *et al*, 2007b). Only hits above 60 bits and whose scores lied within 10% of the best score were considered.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Chris Creevey and the other Bork laboratory members for their valuable input, as well as four anonymous reviewers for their constructive remarks. This research is supported by the European Union FP7 Program Contract no. HEALTH-F4–2007-201052.

Author contributions: JR and PB designed the research. JR, IL, TY, LJJ analyzed the data and/or provided data analysis expertise. JR and PB wrote the paper.

Conflict of interest

The authors declare that they have no conflict of interest.

© 2011 EMBO and Macmillan Publishers Limited

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403–410

Arrigo KR (2005) Marine microorganisms and global nutrient cycles. *Nature* **437:** 349–355

Baas Becking LGM (1934) *Geobiologie of inleiding tot de milieukunde*. The Hague, The Netherlands: W.P. Van Stockum & Zoon

Behrenfeld MJ, O'Malley RT, Siegel DA, McClain CR, Sarmiento JL, Feldman GC, Milligan AJ, Falkowski PG, Letelier RM, Boss ES (2006) Climate-driven trends in contemporary ocean productivity. *Nature* **444:** 752–755

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B* **57:** 289–300

Biers EJ, Sun S, Howard EC (2009) Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Appl Environ Microbiol* **75:** 2221–2229

Boyer TP, Antonov JI, Garcia HE, Johnson DR, Locarnini RA, Mishonov AV, Pitcher MT, Baranova OK, Smolyar IV (2006) *World Ocean Database 2005*. Washington, DC: US Government Printing Office

Chao A (1984) Non-parametric estimation of the number of classes in a population. *Scand J Stat* **11:** 265–270

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* **311:** 1283–1287

Colinvaux PA (1973) *Introduction to Ecology*. New York: Wiley

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311:** 496–503

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R *et al* (2008) Functional metagenomic profiling of nine biomes. *Nature* **452:** 629–632

Falkowski PG, Barber RT, Smetacek VV (1998) Biogeochemical controls and feedbacks on ocean primary production. *Science* **281:** 200–207

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281:** 237–240

Finlay BJ (2002) Global dispersal of free-living microbial eukaryote species. *Science* **296:** 1061–1063

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP (2009) The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323:** 741–746

Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, Green JL, Brown JH (2008) A latitudinal diversity gradient in planktonic marine bacteria. *Proc Natl Acad Sci USA* **105:** 7774–7778

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL, Swings J (2005) Opinion: re-evaluating prokaryotic species. *Nat Rev Microbiol* **3:** 733–739

Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, Letunic I, Yamada T, Paccanaro A, Jensen LJ, Snyder M, Bork P, Gerstein MB (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci USA* **106:** 1374–1379

Gilbert JA, Field D, Swift P, Newbold L, Oliver A, Smyth T, Somerfield PJ, Huse S, Joint I (2009) The seasonal structure of microbial communities in the Western English Channel. *Environ Microbiol* 2009 **11:** 3132–3139

Goslee SC, Urban DL (2007) The ecodist package for dissimilarity-based analysis of ecological data. *J Stat Software* **22:** i07

Green JL, Bohannan BJ, Whitaker RJ (2008) Microbial biogeography: from taxonomy to traits. *Science* **320:** 1039–1043

Grieser J (2002) *Local Climate estimator 1.0. United Nations Food and Agriculture Organisation*. Rome, Italy: Environment and Natural Resources Service

Harrington ED, Singh AH, Doerks T, Letunic I, von Mering C, Jensen LJ, Raes J, Bork P (2007) Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proc Natl Acad Sci USA* **104:** 13913–13918

Hotelling H (1936) Relations between two sets of variants. *Biometrika* **28:** 321–377

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320:** 1081–1085

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* **17:** 377–386

Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, Chisholm SW (2006) Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science* **311:** 1737–1740

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36:** D480–D484

Key RM, Kozyr A, Sabine CL, Lee K, Wanninkhof R, Bullister J, Feely RA, Millero F, Mordy C, Peng T-H (2004) A global ocean carbon climatology: results from GLODAP. *Global Biogeochem Cycles* **18:** GB4031

Kondoh M (2001) Unifying the relationships of species richness to productivity and disturbance. *Proc Biol Sci* **268:** 269–271

Krebs CJ (2001) *Ecology: The Experimental Analysis of Distribution and Abundance*. San Francisco, London: Benjamin-Cummings

Kunin V, Raes J, Harris JK, Spear JR, Walker JJ, Ivanova N, Von Mering C, Bebout BM, Pace NR, Bork P, Hugenholtz P (2008) Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. *Mol Syst Biol* **4:** 198

Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S, DeMaere MZ, Ting L, Ertan H, Johnson J, Ferriera S, Lapidus A, Anderson I, Kyrpides N, Munk AC, Detter C, Han CS, Brown MV, Robb FT, Kjelleberg S *et al* (2009) The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci USA* **106:** 15527–15533

Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23:** 127–128

Lindeman RL (1942) Experimental simulation of winter anaerobiosis in a senescent lake. *Ecology* **23:** 1–13

Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res* **34:** D332–D334

Loreau M, Naeem S, Inchausti P, Bengtsson J, Grime JP, Hector A, Hooper DU, Huston MA, Raffaelli D, Schmid B, Tilman D, Wardle DA (2001) Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science* **294:** 804–808

Lozupone CA, Knight R (2007) Global patterns in bacterial diversity. *Proc Natl Acad Sci USA* **104:** 11436–11440

Magurran AE (1988) *Ecological Diversity and its Measurement*. Princeton, NJ: Princeton University Press

Mahenthiralingam E, Baldwin A, Drevinek P, Vanlaere E, Vandamme P, LiPuma JJ, Dowson CG (2006) Multilocus sequence typing breathes life into a microbial metagenome. *PLoS ONE* **1:** e17

Martiny JB, Bohannan BJ, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR, Morin PJ, Naeem S, Ovreas L, Reysenbach AL, Smith VH, Staley JT (2006) Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4:** 102–112

McGill BJ, Enquist BJ, Weiher E, Westoby M (2006) Rebuilding community ecology from functional traits. *Trends Ecol Evol* **21:** 178–185

Meyer B, Kuever J (2007) Molecular analysis of the distribution and phylogeny of dissimilatory adenosine-5′-phosphosulfate reductase-encoding genes (aprBA) among sulfur-oxidizing prokaryotes. *Microbiology* **153:** 3478–3498

© 2011 EMBO and Macmillan Publishers Limited

Monterey G, Levitus S (1997) *Seasonal Variability of Mixed Layer Depth for the World Ocean*. Washington, DC: US Government Printing Office

Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Solymos P, Stevens H, Wagner H (2008) Vegan: Community Ecology Package. *R package version* **1:** 15–11

Patel PV, Gianoulis TA, Bjornson RD, Yip KY, Engelman DM, Gerstein MB (2010) Analysis of membrane proteins in metagenomics: networks of correlated environmental features and protein families. *Genome Res* **20:** 960–971

Pommier T, Canback B, Riemann L, Bostrom KH, Simu K, Lundberg P, Tunlid A, Hagstrom A (2007) Global patterns of diversity and community structure in marine bacterioplankton. *Mol Ecol* **16:** 867–880

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y *et al* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464:** 59–65

Raes J, Bork P (2008) Molecular eco-systems biology: towards an understanding of community function. *Nat Rev Microbiol* **6:** 693–699

Raes J, Foerstner KU, Bork P (2007) Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol* **10:** 490–498

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K *et al* (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5:** e77

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: a community resource for metagenomics. *PLoS Biol* **5:** e75

Telford RJ, Vandvik V, Birks HJ (2006) Dispersal limitations matter for microbial morphospecies. *Science* **312:** 1015

Tilman D, Knops J, Wedin D, Reich P, Ritchie M, Siemann E (1997) The influence of functional diversity and composition on ecosystem processes. *Science* **277:** 1300–1302

Tringe SG, Rubin EM (2005a) Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* **6:** 805–814

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM (2005b) Comparative metagenomics of microbial communities. *Science* **308:** 554–557

Tripp HJ, Kitner JB, Schwalbach MS, Dacey JW, Wilhelm LJ, Giovannoni SJ (2008) SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature* **452:** 741–744

Vaulot D, Marie D, Olson RJ, Chisholm SW (1995) Growth of Prochlorococcus, a photosynthetic prokaryote, in the Equatorial Pacific Ocean. *Science* **268:** 1480–1482

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P (2007a) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315:** 1126–1130

von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P (2007b) STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* **35:** D358–D362

von Mering C, Zdobnov EM, Tsoka S, Ciccarelli FD, Pereira-Leal JB, Ouzounis CA, Bork P (2003) Genome evolution reveals biochemical networks and functional modules. *Proc Natl Acad Sci USA* **100:** 15428–15433

Woebken D, Teeling H, Wecker P, Dumitriu A, Kostadinov I, Delong EF, Amann R, Glockner FO (2007) Fosmids of novel marine Planctomycetes from the Namibian and Oregon coast upwelling systems and their cross-comparison with planctomycete genomes. *ISME J* **1:** 419–435

Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. *PLoS Comput Biol* **6:** e1000667

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia JM, Soergel DA, Zhai Y *et al* (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* **5:** e16

SOME RIGHTS RESERVED *Molecular Systems Biology* is an open-access journal published by *European Molecular Biology Organization* and *Nature Publishing Group*. This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported License.

© 2011 EMBO and Macmillan Publishers Limited