Supplementary File 4: Methods Summary and Supplementary Methods

Methods Summary

Constructing and implementing an SDM model requires local measurements of community composition and rasters of environmental data. For measurements of community composition, we assembled a database of rDNA data from 377 marine samples from 164 distinct locations with depth <150 m (Supplementary Figure S1). From this database, we excluded samples from vents, anoxic water, sediment, and fresh water. Data came from four sources (Supplementary Table S1); we used MICROBIS for our primary analysis and the other sources for model selection and validation. Three data sources contributed 16S sequences, and reference [3] contributed ARISA data (Supplementary Table S1). OTUs for the ARISA data were obtained from reference [3]. For the non-ARISA data, we considered OTUs defined by reference-based classification and *de novo* clustering.

For the rasters of environmental data, from 45 environmental variables mapped at a 0.5° latitude/longitude resolution across the world ocean, we selected 21 (starred in Supplementary Table S2) that correlated with diversity, were not highly correlated with each other (Supplementary Table S3), and had multivariate environmental similarity surface scores (MESS; [16]) greater than -20 for 99.5% of the world ocean (Table S4). Incorporation of MESS scores ensured that models could be projected into geographic space with minimal extrapolation [16]. Many of the rasters were depth- and month-specific, although these were often less predictive than their averaged counterparts.

To construct SDMs, we fit models using the MICROBIS data, and performed extensive variable selection and validation analyses using all four data sets. We constructed SDMs with linear or non-linear models, rarefaction depths of 4266 or 150 rDNA sequences per sample (with more than 4266 sequences, many samples would have to be excluded), and sequences classified using *de novo* clustering or reference-based classification. Regardless of the methodology, the resulting maps showed temperate diversity peaks in the winter (Figure 1 and Supplementary Figures S2-S5). Thus, we focus on a linear model at a rarefaction depth of 4266 sequences, with *de novo* sequence clustering. To estimate ranges of individual taxa, we used SDMs with a logistic regression model [17].

Supplementary Methods

Reference database-based sequence clustering

The methods used for reference-based sequence classification of the reference [2] and [3] data are described in the respective publications. We used the diversity measurements that resulted from the application of these procedures in these studies for our analyses. For the Global Ocean Survey (GOS) and MICROBIS data, to classify 16S sequences into taxonomic groups (e.g., genera), we first downloaded the 13,550 Global Ocean Survey 16S sequences from CAMERA (http://camera.calit2.net) and the 7,466,321 MICROBIS 16S sequences from the MICROBIS website (http://icomm.mbl.edu/microbis). We then annotated each 16S sequence using the command-line version of the Ribosomal Database Project's 16S sequence classifier (RDP Classifier v2.3, http://sourceforge.net/projects/rdp-classifier; [19]), which is a naïve Bayesian classifier that is trained on the high-quality 16S database curated by the Ribosomal Database Project (http://rdp.cme.msu.edu). It evaluates each sequence independently, assigns a taxonomy string from domain to genus to each sequence,

and provides a confidence estimate of the classification at each taxonomic level in the form of a bootstrap statistic. The output file was parsed using in-house Java scripts that extract a taxonomic annotation for each sequence. We used all taxonomic annotations regardless of their bootstrap values: introducing a bootstrap threshold (e.g., 50%) introduces significant bias to the data set because sequences with high similarity to known bacterial genera are not evenly distributed across latitudes. After rarefaction, the relative abundance of each taxonomic group was calculated by dividing the number of sequences assigned to the group by the total number of rarefied sequences.

De novo sequence clustering

We analyzed *de novo* clustered OTUs that were generated in previous studies that used the same data. MICROBIS sequences were clustered following the methods described in reference [18] and clustering annotations were downloaded from The Visualization and Analysis of Microbial Population Structures website (http://vamps.mbl.edu). The methods used for clustering the reference [2] sequences are described in reference [2], and we used the diversity measurements that resulted from the clustering that was done therein. *De novo* clustering these sequences is not practical because they were generated using shotgun sequencing. Likewise, *de novo* clustering is impractical for the reference [3] ARISA data.

References

1. Amaral-Zettler L, Artigas LF, Baross J, Bharathi L, Boetius A, Chandramohan D, et al. (2010). A global census of marine microbes, In: Life in the World's Oceans: Diversity, Distribution

and Abundance. Blackwell Publishing Ltd., Oxford, (Ed. McIntyre), 223-245.

2. Pommier T, Canback B, Riemann L, Bostrom KH, Simu K, Lundberg P et al. (2007). Global patterns of diversity and community structure in marine bacterioplankton. Mol Ecol **16:** 867–880.
3. 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.
4. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S et al. (2007). The Sorcerer II global ocean sampling expedition: northwest atlantic through eastern tropical pacific. PLoS Biol **5:** e77.
5. Garcia HE, Locarnini RA, Boyer TP, Antonov JI, Baranova OK, Zweng MM et al. (2010). World Ocean Atlas 2009, Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation. U.S. Government Printing Office.
6. Tyberghein L, Verbruggen H, Pauly K, Troupin C, Mineur F, Clerck OD. (2012). Bio-oracle: a global environmental dataset for marine species distribution modelling. Global Ecol Biogeogr **21:** 272–281.
7. NASA earth observations. (2012). http://neo.sci.gsfc.nasa.gov/.
8. Stott J. (2012). Earthtools. http://www.earthtools.org/webservices.htm.
9. Ready J, Kaschner K, South AB, Eastwood PD, Rees T, Rius J et al. (2010). Predicting the distributions of marine organisms at the global scale. Ecol Model **221:** 467–478.
10. Jickells TD, An ZS, Andersen KK, Baker AR, Bergametti G, Brooks N et al. (2005). Global iron connections between desert dust, ocean biogeochemistry, and climate. Science **308:** 67–71.

11. Garcia HE, Locarnini RA, Boyer TP, Antonov JI, Baranova OK, Zweng MM et al. (2010). World Ocean Atlas 2009, Volume 4: Nutrients (phosphate, nitrate, silicate). U.S. Government Printing Office.
12. Locarnini RA, Mishonov AV, Antonov JI, Boyer TP, Garcia HE, Baranova OK, et al. (2010). World Ocean Atlas 2009, Volume 1: Temperature. U.S. Government Printing Office.
13. Montegut CB, Madec G, Fischer AS, Lazar A. (2004). Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology. Journal of Geophysical Research **109:** C12003.
14. Antonov JI, Seidov D, Boyer TP, Locarnini RA, Mishonov AV, Garcia HE, et al. (2010). World Ocean Atlas 2009, Volume 2: Salinity. U.S. Government Printing Office.
15. Halpern BS, Walbridge S, Selkoe KA, Kappel CV, Micheli F, D'Agrosa C et al. (2008). A global map of human impact on marine ecosystems. Science **319:** 948–952.
16. Elith J, Kearney M, Phillips S. (2010). The art of modelling range-shifting species. Method Ecol Evol 1: 330–342.
17. Franklin J, Miller JA. (2009). Mapping species distributions: Spatial Inference and Prediction. Cambridge University Press: Cambridge, UK.
18. Huse SM, Welch DM, Morrison HG, Sogin ML. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Env Microbiology **12:** 1889–1898.
19. Wang Q, Garrity G, Tiedje J, Cole J. (2007). Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol **73:** 5261-5267.