Title: Biogeographic patterns in ocean microbes emerge in a neutral agent-based model

Authors: Ferdi L. Hellweger[1]*, Erik van Sebille[2], Neil D. Fredrick[1]

Affiliations:

[1]Civil & Environmental Engineering, Northeastern University, Boston, Massachusetts, USA.

[2]ARC Centre of Excellence for Climate System Science & School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia

*Correspondence to: ferdi@coe.neu.edu

Abstract: A key question in ecology and evolution is the relative role of natural selection and neutral evolution in producing biogeographic patterns. Here we quantify the role of neutral processes by simulating division, mutation and death of 100k individual marine bacteria cells with full 1 Mbp genomes in a global surface ocean circulation model. The model is run for up to 100k years and output is analyzed using BLAST alignment and metagenomics fragment recruitment. Simulations show the production and maintenance of biogeographic patterns, characterized by distinct provinces subject to mixing and periodic takeovers by neighbors (coalescence), after which neutral evolution re-establishes the province and the patterns reorganize. The emergent patterns are substantial (e.g., down to 99.5% DNA identity between North and Central Pacific provinces) and suggest that microbes evolve faster than ocean currents can disperse them. This approach can also be used to explore environmental selection.

Main Text:

An important ongoing endeavor in ecology and evolution is to understand the mechanisms underlying the geographic distribution patterns of organisms. Natural selection by contemporary environmental factors acting on adaptive mutations or a persistent seed bank of species is one mechanism that can create such patterns. Neutral evolution (selectively neutral mutations and genetic drift) coupled with dispersal limitation or isolation is another mechanism (*1-6*). These processes are not mutually exclusive and, for microbes in the global surface ocean, molecular observations (e.g., shotgun sequencing, (*7*)) provide support for the role of both mechanisms (*8-11*).

Here we ask: How does neutral evolution influence the biogeographic distribution of surface ocean microbes? To what extent does dispersion allow for different operational taxonomic units (OTUs) to develop and persist? Are there emerging spatial patterns (e.g., provinces, (*12*)) and how do these change in time?

Several approaches are available to quantify the contribution of the various processes in generating and maintaining biogeographic patterns among ocean microbes (*2*). A common empirical approach involves correlating observations (e.g., microbial composition) with environmental variables, subtracting out this environment effect and then correlating with

geographic distance. In the ocean, hydrodynamic models coupled with tracers, either Eulerian concentration or Lagrangian particles, can be used as a measure of "connectivity" to supplement empirical relations (*10, 13*). A problem with this approach is that one can never be sure that the distance effect is not actually caused by an unmeasured environmental variable. Mechanistic models constitute an alternative approach. Eulerian models with coupled phytoplankton ecology and biogeochemistry can simulate biogeographic patterns produced by environmental selection (*14*). However, the Eulerian approach generally assumes all species are present everywhere and does not consider dispersal limitation.

An alternative approach is to use a neutral model, where environmental selection is excluded. Such a model can be used as a null hypothesis, where deviation from observed patterns can help quantify the effect of the omitted factors (e.g., selection) (*6*, *15, 16*). A neutral model can also be used to directly predict biogeographic patterns resulting from neutral processes. Of course, this is only meaningful if the model parameters are sufficiently constrained. For bacteria in the open ocean, hydrodynamic transport is relatively well-understood and rates of growth and base pair mutation are available (*17-19*), allowing us to make meaningful predictions without prior calibration (e.g., dispersal parameter, (*15*)).

We model individual bacteria using an agent- (i.e., Lagrangian-, individual-) based approach (*20-24*). The large number of ocean microbes, one estimate is 3.6×10^{28} prokaryotes in the surface ocean (*25*), makes simulating all individuals impossible and we therefore model a smaller number of representative cells, referred to as super individuals (*21, 24*). This prohibits us from explicitly simulating absolute diversity or species richness. However, biogeographic distribution quantified as nucleotide divergence between cells at different locations (total nucleotide difference of cells from different locations minus local nucleotide diversity) (*26*) is independent of population size. This is supported by theory (*27*) and an application of our full model to a simple two box system (*28*), and it can be understood as follows. Consider two locations (provinces) with equal-sized populations and limited mixing. Cells continuously migrate between these provinces and periodically one succeeds to take over the recipient province (coalescence). At that time the nucleotide divergence between the subpopulations is eliminated and then it increases in linear, clockwise manner that is independent of the population size until the next coalescence event. Thus, the frequency of coalescence events determines the average divergence. The rate at which cells migrate between provinces is proportional to the population size, but the probability of a migrant cell taking over the population is inversely proportional to the population size. As a result the frequency of coalescence (migration and takeover) is independent of the population size. Despite the large scaling in population size (model versus reality), the super individual approach should produce meaningful estimates of nucleotide divergence.

Our model simulates a population of individual cells. The cells die at a constant rate and divide at a rate that depends on the local cell concentration using a logistic approach. Death and division are stochastic. As the local cell density approaches a specified carrying capacity, the growth rate slows down. The parameters were adjusted to produce reasonable growth rates (0.14 day^{-1} (*29*)) and a population of about 100k cells distributed at a relatively constant density. Cells have a 1 Mbp genome (i.e., an array of A, T, C and G letters) subject to mutation, where the mutation rate was increased from the point mutation rate (i.e., $m = 5.4\times10^{-10}$ bp/division for

Escherichia coli, (*17*)) to account for base pair changes by recombination (r / m = 63.1 for *Pelagibacter ubique*, (*18, 19*), $m + r = 3.5\times10^{-8}$ bp/division). Cells are transported in a Lagrangian manner horizontally within the surface layer using a 1/10º eddy-resolving circulation model, spanning the oceans from 75ºS to 75ºN (*30*), with cycling of a 31-year period to allow for simulations up to 100k years (*28*).

We first consider the diversity in the model, which is not a realistic estimate of the actual diversity in the surface ocean microbe population because we simulate super-individuals, but it illustrates the behavior of model. When the model starts with a diverse population (each cell has a different genome) and no mutation, diversity decreases monotonically as OTUs are lost by extinction and not gained by mutation (Fig. 1A). After about 100 years, the population consists of about 10 OTUs, resident in relatively distinct spatial regions and the rate of OTU loss becomes limited by dispersal between these provinces (*28*). At that time, the model starts to predict higher OTU richness than a neutral theory model that does not consider dispersal limitation (*31*). The rate of OTU loss becomes low, but the populations continue to mix (*28*) and the probability of extinction remains greater than zero. At 100k years the model includes two OTUs, in the Southern Ocean and everywhere else. The model should eventually reduce to one OTU, although this is not realized in the 100k year simulation.

When the model is initialized with a diverse population and includes mutation it also exhibits an initial rapid loss in diversity, but then levels off at an OTU richness slightly higher than the simulation without mutation (Fig. 1B). For these simulations the diversity is determined from a sample of the population (100 cells) by performing pairwise whole-genome BLAST alignment, identifying OTUs using 99.9% whole genome identity cutoff and then up-scaling to the true richness (in the model) using Chao1, a non-parametric species estimator that extrapolates from the sample data to "true" richness (see Methods). The OTU richness is variable over time because of stochastic transport and sampling (see Fig. S4), but that is identical for all simulations. Therefore, the difference in OTU richness between the simulations with and without mutation can be attributed solely to mutation. The difference is relatively small, but increases with higher taxonomic resolution (99.95% cutoff) or mutation rate (X3). For a simulation starting with a uniform population (all cells have the same genome) and including mutation, the OTU richness starts at one and then increases once sufficient mutations accumulate to exceed the OTU threshold. After about 200 years, the simulations starting diverse and uniform converge. At that time model has reached a dynamic steady state where the rate of OTU loss by extinction is balanced by the rate of OTU gain by mutation. From a practical perspective, this shows that specifying different initial conditions or running the model any longer would not change the diversity.

The model is then used to explore the role of neutral evolution in producing biogeographic patterns. As an example, we compared the genomes of cells from Hawaii and the Gulf of Alaska (Fig. 2B). For the simulation starting diverse without mutation, the difference (nucleotide divergence) is 100% until about 700 years when it abruptly decreases to 0% (Fig. 2A). This is caused by a takeover of the Central Pacific province by a cell from the North Pacific province, or a coalescence of these two subpopulations (*27*) (see also Movie S1 around 700 years). The simulation starting diverse with mutation also starts at 100% and decreases at the coalescence event, but then increases again as the two subpopulations diverge. The simulation starting

uniform initially has 0% difference, but immediately starts to increase and then converges with the simulation starting diverse and including mutation. Coalescence events are stochastic (see also Fig. S4B) and over the 1,500 year simulation period we observed two. There are also occasional abrupt drops in nucleotide divergence, which are due to vagrant cells that enter a province, but do not establish. The magnitude of nucleotide divergence is a function of the growth and mutation rates (Figs. S5 and S6).

The time between coalescence events puts a limit to how much two provinces can diverge and in this case the model predicts up to 0.5% difference (99.5% identity). These results can be related to observations. If two cells are sampled from these two locations and their genomes are sequenced and compared, 0-0.5% of the observed difference can be attributed to neutral processes. This level of divergence is substantial, but considerably lower than what is commonly considered a species (>95% identity). We map out the biogeographic pattern produced by neutral evolution for Hawaii compared with all locations across the globe (Fig. 2B). The model predicts that nucleotide divergence generally increases with distance from Hawaii. However, the divergence is larger for the North Pacific than the Indian Ocean, so distance and/or the presence of landmasses are not necessarily good proxies of dispersal barriers. We also compile this information for all locations compared with all locations into an atlas of neutral biogeography (Table S1).

Finally, we mapped out the biogeographic pattern produced by neutral evolution using fragment recruitment (*7*), which is akin to *in silico* DNA hybridization. Specifically, we take the single-cell genomes (SCGs) of the OTUs that were maintained in the simulation starting uniform at 1,400 years (see Fig. 1B) and recruit fragments collected on a 10×10º grid. We assigned each grid box to the highest-recruiting SCG (i.e., dominant OTU) and color them accordingly, which illustrated the provinces produced by neutral evolution and maintained by dispersal limitation (Fig. 3). The resulting map is a function of the taxonomic cutoff used and at less than 99% the model does not show any diversity or biogeographic patterns. This is expected, because over 1,400 years the genomes only evolve by 0.5%, corresponding to a 99% identity. However, at this time, there are still provinces that have not yet experienced a takeover or coalescence event (Fig. 1A), and those are expected to continue to diverge beyond the 99% threshold for longer times. This distribution can change temporarily as a result of takeover events. For example, at times the Central Pacific and North Pacific provinces were distinguishable at 99.5% (0.5% difference, see Fig. 2A).

We conclude that neutral evolution (neutral mutations and genetic drift) coupled with dispersal limitation can produce substantial biogeographic patterns in the global surface ocean microbe population. Microbes evolve faster than the ocean circulation can disperse them, a feature that can also be seen in molecular observations (*10*). The patterns are dynamic. Provinces gradually emerge as subpopulations diverge by neutral evolution and periodically collapse due to coalescence. Neutral processes, along with environmental selection, have to be considered in future research on microbial biogeography and our results provide a quantitative benchmark for their potential role. Our results contrast the notion that “everything is everywhere” (*11, 32*) and may have important implications for how the oceans will respond to global change. Our model provides insights into the role of neutral evolution in shaping biogeographic patterns. The biology in the model is relatively simple and future work may build on this by considering more

spatial and temporal patterns (e.g., carrying capacity, division and death rates based on ocean productivity) and more explicit representation of processes (e.g., recombination). Our modeling approach can theoretically be used to explore environmental selection as well. This will require relating genes to function, which is difficult but can be done for select genes or at the genome level (*23*, *33*).

References and Notes:

1. A. Ramette, J. M. Tiedje, Multiscale responses of microbial life to spatial distance and environmental heterogeneity in a patchy ecosystem. *Proceedings of the National Academy of Sciences* **104**, 2761 (2007).
2. C. A. Hanson, J. A. Fuhrman, M. C. Horner-Devine, J. B. H. Martiny, Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat Rev Micro* **10**, 497 (2012).
3. R. T. Papke, D. M. Ward, The importance of physical isolation to microbial diversification. *FEMS Microbiology Ecology* **48**, 293 (2004).
4. M. V. Brown, M. Ostrowski, J. J. Grzymski, F. M. Lauro, A trait based perspective on the biogeography of common and abundant marine bacterioplankton clades. *Marine Genomics* **15**, 17 (2014).
5. R. H. MacArthur, E. O. Wilson, *The Theory of Island Biogeography* (Princeton Univ. Press, Princeton, NJ, 1967).
6. S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton Univ. Press, Princeton, NJ, 2001).
7. D. B. Rusch *et al.*, The <italic>Sorcerer II</italic> Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**, e77 (2007).
8. B. K. Swan *et al.*, Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proceedings of the National Academy of Sciences* **110**, 11463 (2013).
9. J.-F. Ghiglione *et al.*, Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proceedings of the National Academy of Sciences* **109**, 17633 (2012).
10. A. C. Martiny, A. P. K. Tai, D. Veneziano, F. Primeau, S. W. Chisholm, Taxonomic resolution, ecotypes and the biogeography of Prochlorococcus. *Environmental Microbiology* **11**, 823 (2009).
11. S. M. Gibbons *et al.*, Evidence for a persistent microbial seed bank throughout the global ocean. *Proceedings of the National Academy of Sciences* **110**, 4651 (2013).
12. J. B. H. Martiny *et al.*, Microbial biogeography: putting microorganisms on the map. *Nat Rev Micro* **4**, 102 (2006).
13. D. Wilkins, E. van Sebille, S. R. Rintoul, F. M. Lauro, R. Cavicchioli, Advection shapes Southern Ocean microbial assemblages independent of distance and environment effects. *Nat Commun* **4**, 2457 (2013).
14. A. D. Barton, S. Dutkiewicz, G. Flierl, J. Bragg, M. J. Follows, Patterns of Diversity in Marine Phytoplankton. *Science* **327**, 1509 (2010).
15. R. Condit *et al.*, Beta-Diversity in Tropical Forest Trees. *Science* **295**, 666 (2002).
16. W. J. Sul, T. A. Oliver, H. W. Ducklow, L. A. Amaral-Zettler, M. L. Sogin, Marine bacteria exhibit a bipolar distribution. *Proceedings of the National Academy of Sciences* **110**, 2342 (2013).
17. J. W. Drake, B. Charlesworth, D. Charlesworth, J. F. Crow, Rates of Spontaneous Mutation. *Genetics* **148**, 1667 (1998).
18. K. L. Vergin *et al.*, High intraspecific recombination rate in a native population of Candidatus Pelagibacter ubique (SAR11). *Environmental Microbiology* **9**, 2430 (2007).

19. M. Vos, X. Didelot, A comparison of homologous recombination rates in bacteria and archaea. *ISME J* **3**, 199 (2008).

20. J.-U. Kreft *et al.*, Mighty small: Observing and modeling individual microbes becomes big science. *Proceedings of the National Academy of Sciences* **110**, 18027 (2013).

21. F. L. Hellweger, V. Bucci, A bunch of tiny individuals—Individual-based modeling for microbes. *Ecological Modelling* **220**, 8 (2009).

22. P. G. Falkowski, C. D. Wirick, A simulation model of the effects of vertical mixing on primary productivity. *Marine Biology* **65**, 69 (1981).

23. F. L. Hellweger, Carrying photosynthesis genes increases ecological fitness of cyanophage *in silico*. *Environmental Microbiology* **11**, 1386 (2009).

24. J. R. Clark, T. M. Lenton, H. T. Williams, S. J. Daines, Environmental selection and resource allocation determine spatial patterns in picophytoplankton cell size. *Limnol. Oceanogr* **58**, 1008 (2013).

25. W. B. Whitman, D. C. Coleman, W. J. Wiebe, Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences* **95**, 6578 (1998).

26. M. Nei, W. H. Li, Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences* **76**, 5269 (1979).

27. M. Slatkin, Isolation by Distance in Equilibrium and Non-Equilibrium Populations. *Evolution* **47**, 264 (1993).

28. Materials and methods are available as supplementary materials on Science Online.

29. H. Ducklow, Bacterial production and biomass in the oceans. *Microbial ecology of the oceans* **1**, 85 (2000).

30. Y. Masumoto *et al.*, A fifty-year eddy-resolving simulation of the world ocean: Preliminary outcomes of OFES (OGCM for the Earth Simulator). *J. Earth Simulator* **1**, 35 (2004).

31. J. M. Halley, Y. Iwasa, Neutral theory as a predictor of avifaunal extinctions after habitat loss. *Proceedings of the National Academy of Sciences* **108**, 2316 (2011).

32. P. Cermeño, P. G. Falkowski, Controls on Diatom Biogeography in the Ocean. *Science* **325**, 1539 (2009).

33. A. M. Feist, M. J. Herrgård, I. Thiele, J. L. Reed, B. Ø. Palsson, Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology* **7**, 129 (2008).

34. S. J. Giovannoni *et al.*, Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science* **309**, 1242 (2005).

35. H. Sasaki *et al.*, in *High Resolution Numerical Modelling of the Atmosphere and Ocean,* K. Hamilton, W. Ohfuchi, Eds. (Springer New York, 2008), pp. 157-185.

36. X. Qin, E. van Sebille, A. Sen Gupta, Quantification of errors induced by temporal resolution on Lagrangian particles in an eddy-resolving model. *Ocean Modelling* **76**, 20 (2014).

37. C. B. Paris, J. Helgers, E. Van Sebille, A. Srinivasan, Connectivity Modeling System: A probabilistic modeling tool for the multi-scale tracking of biotic and abiotic variability in the ocean. *Environmental Modelling & Software* **42**, 47 (2013).

38. H. Luo, M. Csűros, A. L. Hughes, M. A. Moran, Evolution of Divergent Life History Strategies in Marine Alphaproteobacteria. *mBio* **4**, e00373-13 (2013).

39. N. Takahata, The coalescent in two partially isolated diffusion populations. *Genetics Research* **52**, 213 (1988).

40. W.-H. Li, Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. *Genetics* **85**, 331 (1977).

41. S. S. Drijfhout, E. Maier-Reimer, U. Mikolajewicz, Tracing the conveyor belt in the Hamburg large-scale geostrophic ocean general circulation model. *Journal of Geophysical Research: Oceans* **101**, 22563 (1996).

42. E. van Sebille, L. M. Beal, W. E. Johns, Advective Time Scales of Agulhas Leakage to the North Atlantic in Surface Drifter Observations and the 3D OFES Model. *Journal of Physical Oceanography* **41**, 1026 (2011).

43. J. B. Sallée, K. Speer, R. Morrow, R. Lumpkin, An estimate of Lagrangian eddy statistics and diffusion in the mixed layer of the Southern Ocean. *Journal of Marine Research* **66**, 441 (2008).

Acknowledgments: We would like to thank Ramunas Stepanauskas for interesting discussion that inspired this research. Three anonymous reviewers provided constructive criticism. FLH and NDF are supported by grants from the National Science Foundation and NOAA MIT SeaGrant. EvS is supported by the Australian Research Council via grant DE130101336.

Fig. 1. Diversity (OTU richness) in global surface ocean microbes predicted by a neutral agent-based model. Results from several simulations are presented: "Start Diverse": All initial cells have an individual, completely random genome. "Start Uniform": All initial cells have the same, completely random genome. “No Mut.” indicates zero and “X3 Mut.” indicates three times higher mutation rate. “99.95%” indicates 99.95% cutoff (vs. 99.9% used in other analyses). “Theoretical”: Based on neutral theory, not considering dispersal limitation (*31*). (**A**&**B**) OTU richness over time. Note x- and y-axis scales. Start diverse (red) and start uniform (blue) lines overlap after 200 years. (**C**) Life history of an individual cell isolated near Bermuda at time 1,000 years. Color changes demarcate mutation events. This cell was initialized with the *P. ubique* HTCC1062 genome (*34*).

Fig. 2. Biogeographic pattern (nucleotide divergence) in global surface ocean microbes predicted by a neutral agent-based model and quantified by genome alignment. (**A**) Cells from Hawaii (HOT) compared to cells from Gulf of Alaska (GOA) over time. See panel B for locations. See caption Fig. 1 for description of simulations. SCGs collected at GOA (n = 10) were aligned against SCGs collected at HOT (n = 10). 1 Mbp genome size. Start diverse, no mutation (green) and start diverse (red) lines overlap up to 700 years. Start diverse (red) and start uniform (blue) lines overlap after 700 years. (**B**) Cells from Hawaii (HOT) compared to cells from all locations. SCGs collected on a 10×10º grid (n = 5 at each box) were aligned to SCGs from HOT (n = 5). 100 Kbp genome size. “Starting Uniform” simulation (all initial cells have the same, completely random genome). Average of nucleotide divergence over 1,500 years.

Fig. 3. Biogeographic patterns (OTU provinces) in global surface ocean microbes predicted by a neutral agent-based model and quantified by metagenomics fragment recruitment. Alignment of fragments collected on a 10×10º grid (n = 10,000 at each box, l = 1,000 bp) with SCGs from OTUs remaining at 1,400 years (see Fig. 1B). 99.9% and 99.5% BLAST identity. “Starting Uniform” simulation (all initial cells have the same, completely random genome). The colors demarcate areas with common dominant OTU.

Supplementary Materials:

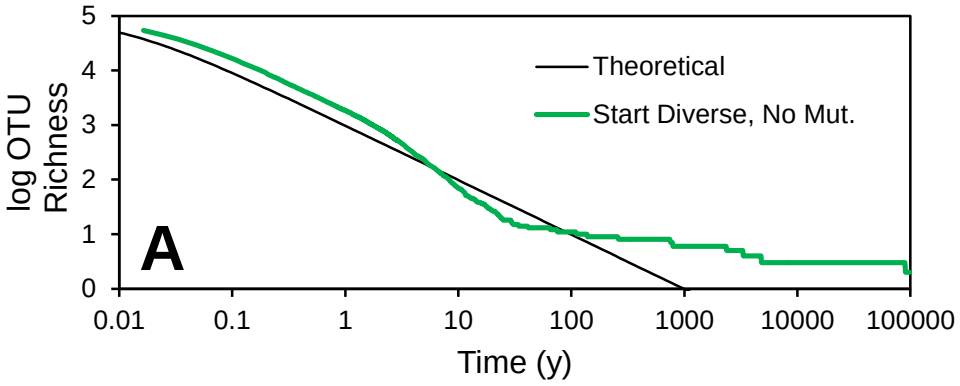
Materials and Methods

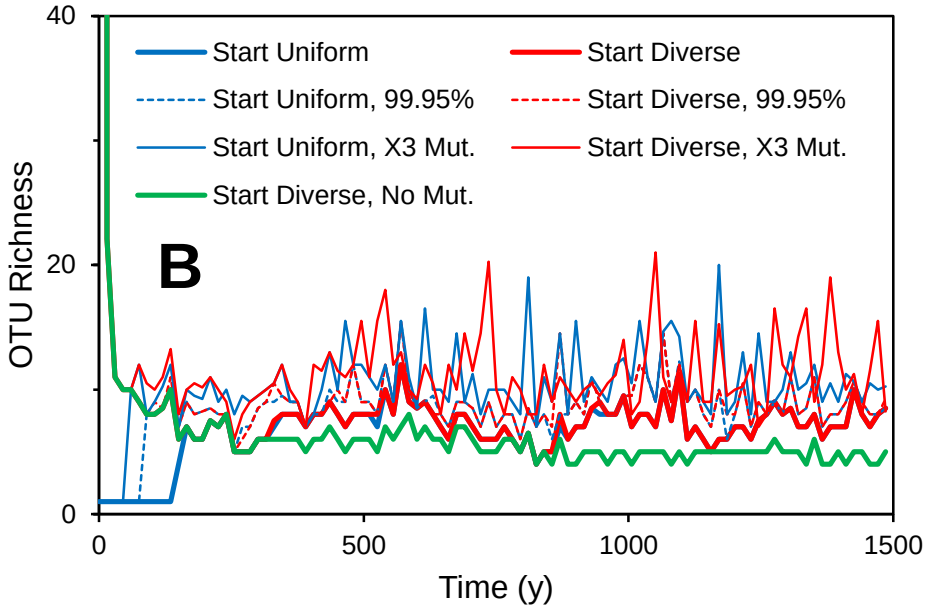
Supplementary Text

Figures S1-S9

Table S1

Movie S1


log OTU Richness
5
4
3
2
1
0
A
Theoretical
Start Diverse, No Mut.
0.01
0.1
1
10
100
1000
10000
100000
Time (y)


OTU Richness
40
20
0
B
Start Uniform
Start Diverse
Start Uniform, 99.95%
Start Diverse, 99.95%
Start Uniform, X3 Mut.
Start Diverse, X3 Mut.
Start Diverse, No Mut.
0
500
1000
1500
Time (y)

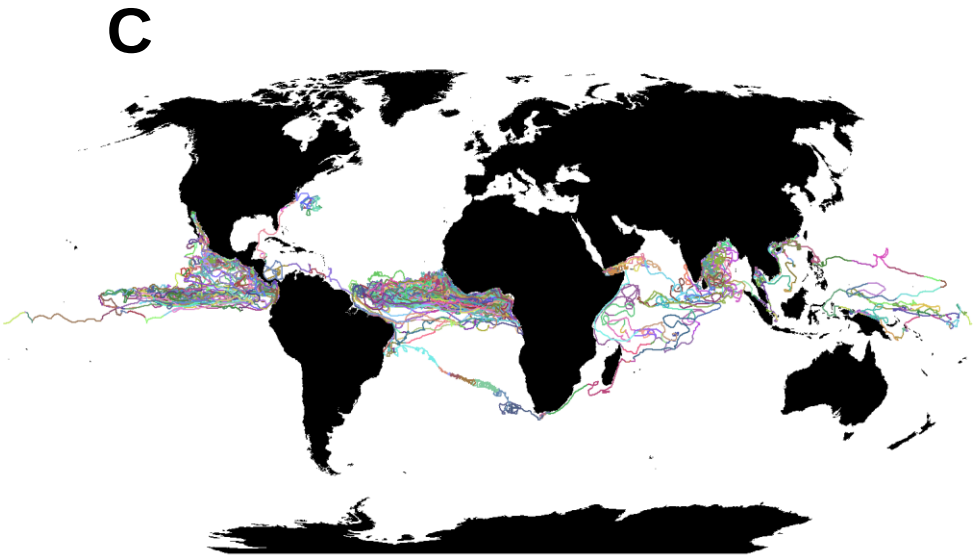

C

Fig. 1

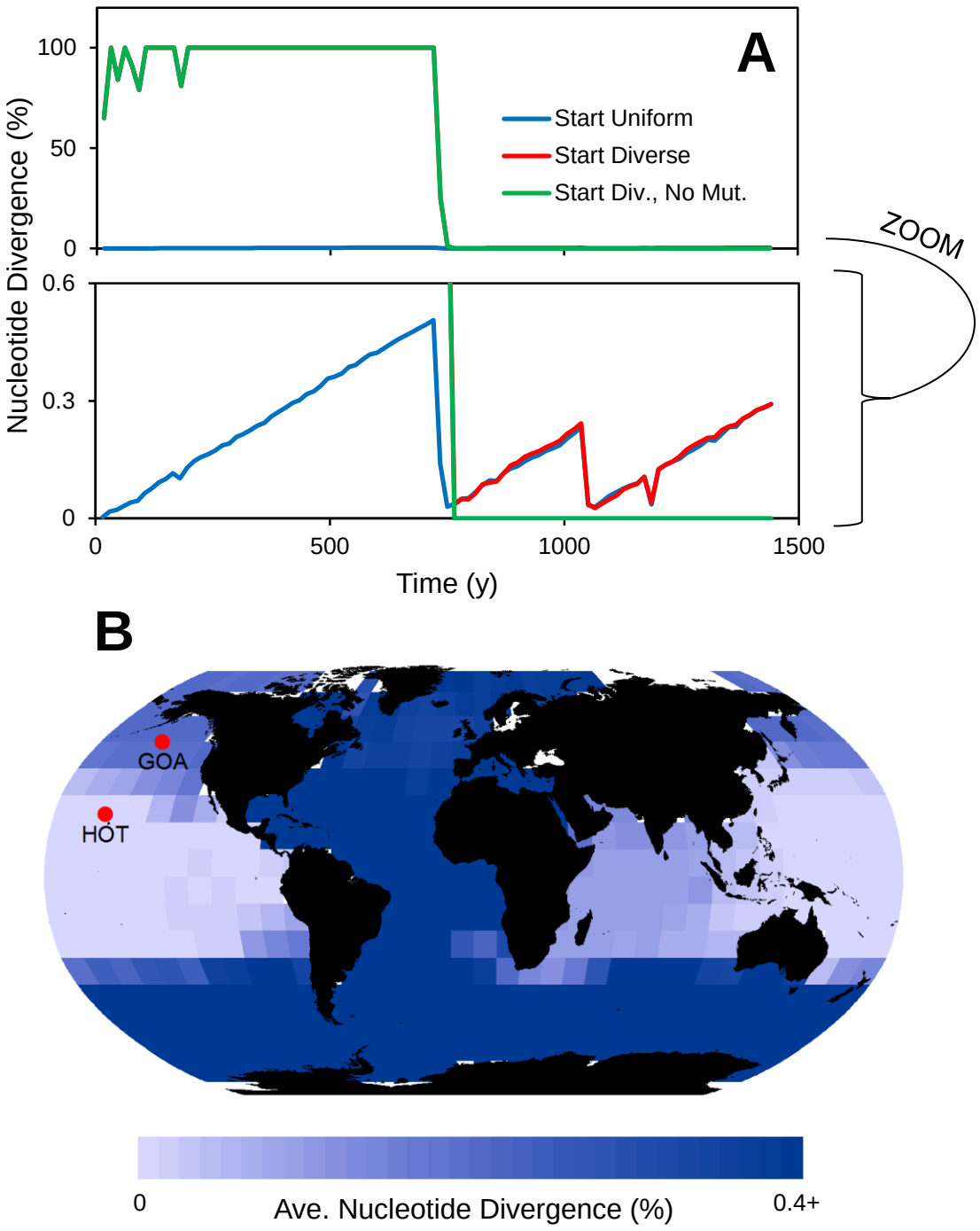

A
100
50
0
Nucleotide Divergence (%)
Start Uniform
Start Diverse
Start Div., No Mut.
ZOOM
0.6
0.3
0
0
500
1000
1500
Time (y)
B
GOA
HOT
0
Ave. Nucleotide Divergence (%)
0.4+

Fig. 2

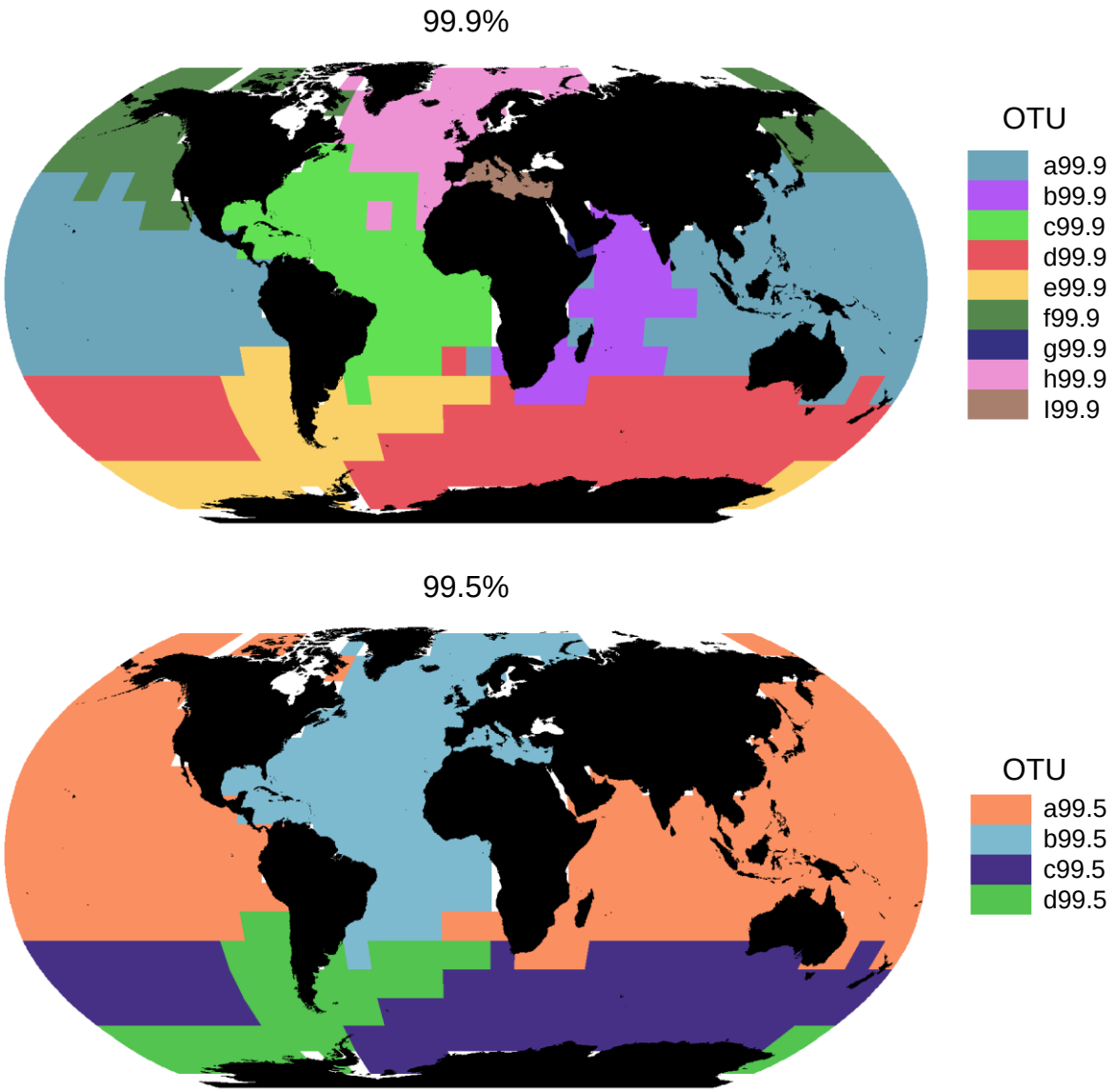

99.9%
OTU
a99.9
b99.9
c99.9
d99.9
e99.9
f99.9
g99.9
h99.9
I99.9
99.5%
OTU
a99.5
b99.5
c99.5
d99.5

Fig. 3