Supplementary Materials for

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

Ferdi L. Hellweger, Erik van Sebille, Neil D. Fredrick

correspondence to: ferdi@coe.neu.edu

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S9
Captions for Table S1
Captions for Movie S1

Other Supplementary Materials for this manuscript includes the following:

Table S1
Movies S1

Materials and Methods

Agent-based modeling of individual bacteria in the global surface ocean

The model simulates individual bacterial cells that divide, die and mutate. The population dynamics are controlled using a logistic approach. That is, cells divide based on a growth rate (k_g, day^{-1}) that is a function of a maximum growth rate ($k_{g,m}$, day^{-1}), and local (5×5º grid box, average dimensions 560×400 km) population size ($P_{x,y}$, cells) and carrying capacity ($K_{x,y}$, cells), according to $k_g = k_{g,m}\,(1–P_{x,y}\,/K_{x,y})$. $K_{x,y}$ varies spatially based on the area of the grid box ($A_{x,y}$, m^2) and a constant areal carrying capacity ($K'_{x,y}$, cells m^{-2}) according to $K_{x,y} = K'_{x,y}\,A_{x,y}$. $P_{x,y}$ varies dynamically during the simulation. A constant death rate (k_d, day^{-1}) is used. Division and death are independent of the cell's genome (i.e., neutral). They are implemented in a stochastic manner (e.g., the probability of a cell dying during a time step is $k_d\,\Delta t$, where Δt (day) is the time step duration). The maximum growth and death rates were assigned to achieve death and division rates consistent with estimates for ocean bacteria ($k_d = 0.5 \times k_{g,m} = 0.14$ day^{-1}, (*29*)). The carrying capacity was adjusted to obtain about 100k cells ($K'_{x,y} = 3.4\times10^{-10}$ cells m^{-2}). Thus, our model simulates about as many cells as there are in a few drops of seawater (5×10^5 cells mL^{-1}, (*25*)).

Each cell has a genome consisting of an array of A, T, C and G letters. For most simulations we use a 1 Mbp, completely random genome. The model can also be run with a real genome, and to illustrate that we perform a simulation with the 1.3 Mbp *P. ubique* (SAR11) HTCC1062 genome (*34*) (Fig. 1C). The genomes are dynamic and subject to change by a simple base pair mutation mechanism, where at division, at a certain frequency, a base pair is randomly selected and changed. The rate was increased from the point mutation rate ($m = 5.4\times10^{-10}$ bp/division, *E. coli*, (*17*)) to account for base pair changes by recombination ($r / m = 63.1$, *P. ubique*, (*18, 19*)), $m + r = 3.5\times10^{-8}$ bp/division). Thus, the model does not explicitly simulate recombination (i.e., copy DNA from one model cell to another), but it accounts for the changes to the genome by this process.

The model simulates cells in the surface layer of the global ocean. Cells are advected within the horizontal velocity field from the Ocean model For the Earth Simulator (OFES) based on the MOM3 ocean model (*30, 35*). OFES is an eddy-resolving ocean model with a horizontal resolution of 1/10º and 54 vertical layers, spanning the oceans from 75ºS to 75ºN and forced by NCEP winds and fluxes. Here, velocity fields from the last 31 years of the simulation (1980-2010) are used. The data are available as 3-day averages, which is sufficient temporal resolution to accurately capture the mesoscale dynamics (*36*). The model cells themselves are advected using the Connectivity Modelling System v1.1 (CMS, (*37*)). In total, 10^5 model cells are released on a regular grid and at 10 m depth. Only the horizontal velocity fields are used and the model cells remain at this depth. The model cells are advected with a time step of 6 hours. An additional horizontal diffusion of 100 m^2/s is used to mimic the mixing by sub-mesoscale processes not resolved within the hydrodynamic model.

To allow for longer simulations the 31-year master history is cycled. Specifically, a continuous long-term (up to 100k years) history is constructed by stitching together pieces with random start/end times from the master history. Individual histories are matched at start/end points in each grid box. Since the number of cells in a box at the start/end times of the histories can be different, this process involves some duplicating and deleting of cells, which causes a loss of diversity, but this occurs infrequently (< 0.1% of actual divisions/deaths). Simulations with and without this cycling are consistent (over the 31 year period, Fig. S1A), and a simulation without mutation loses diversity in agreement with a theoretical estimate (Fig. 1A, (*31*)).

The 100k year simulation time is substantial, but still low in the context of evolution. For example, the occurrence of the common ancestor of *Pelagibacter ubique* strains HTCC1002 and HTCC1062, which have whole genome identity of 98%, is about 5.4M years (*38*). However, the convergence of simulations with different initial conditions (i.e., starting diverse and uniform, Figs. 1B & 2A) suggest that we are running the model for a sufficiently long period so that our results will not change (but see discussion on three provinces remaining at 100k years in Text S2).

Model analysis: Quantifying diversity

We analyze the model output for diversity and biogeographic patterns. For diversity, this is complicated slightly because the model does not explicitly consider species or OTUs. We define OTUs based on whole genome DNA identity, which can be obtained in two ways. For simulations starting with a diverse population and without mutation, the genomes are completely unique and static, and OTUs can be taken as the identification (ID) number of the ancestral cell. That is, the ID number of the initial cell is tracked for each cell and used to identify the OTU. For simulations with mutation (and dynamic genomes) the OTU of a cell cannot be equated to an ancestor's ID and they are determined by doing pairwise BLAST alignment. Calculating the true OTU richness in the model in this manner would require doing BLAST analyses on $10^5 \times 10^5$ model cells, which is not feasible. Therefore, richness is calculated from a sample of 10^2 model cells using the Chao1 estimator. Specifically, the estimated richness (S_1) is a function of the richness in the sample (S_{obs}) and the number of species with one or two representatives (F_1, F_2) using $S_1 = S_{obs} F_1^2 / (2 F_2)$.

The two methods are compared for the simulation starting diverse without mutation (Fig. S1B), where the "true" diversity is determined directly from all individuals in the population and the "estimated" diversity is determined from a sample of the population using BLAST and Chao1. This estimated OTU richness is subject to stochastic variability and can show occasional increases. Early in the simulation the estimated OTU richness is less than the true richness, but later in the simulation, the two methods converge.

We present results from an analytical neutral theory model from the literature (*31*). Specifically, the model predicts the number of species ($S(t)$) as a function of time (t) from

the initial number of species ($S(0)$) and the half-life (t_{50}, i.e., $S(t_{50}) = S(0) / 2$) using $S(t) = S(0) / (1 + t / t_{50})$. The half-life is calculated from the number of individuals (N) and generation time (τ) using $t_{50} = \tau N / S(0)$. The number of individuals, initial number of species and generation time are based on the full model.

Model analysis: Quantifying biogeographic patterns

We consider biogeographic patterns to be difference of DNA from different locations and quantify it by DNA alignment or metagenomics fragment recruitment.

DNA alignment

We use a population genetics approach to quantify biogeographic patterns. For this, we quantify diversity as genomic nucleotide diversity (π) (*26*), the mean percentage of nucleotide differences from pairwise comparisons of a sample of cells. We consider the total nucleotide difference between two locations X and Y (π_{XY}) to be comprised of local nucleotide diversity at each location (π_X, π_Y) and nucleotide divergence between the locations (δ_{XY}). Thus, to calculate δ_{XY}, we randomly sample cells from each location (i.e., grid box) (n = 10 each). Then we calculate π_X by performing whole genome BLAST alignment on each cell pair at that location, calculating the difference for each pair (π_{ij} = 100% – BLAST identity) and the sample (π_X = ave[π_{ij}]). We calculate π_Y in the same manner. Then we calculate π_{XY} by performing alignment on each pair of cells from different locations (i.e., all cells from location X vs. all cells from location Y). Finally, we calculate δ_{XY} as the difference ($\delta_{XY} = \pi_{XY} – [\pi_X + \pi_Y] / 2$). This measure is independent of population size (Text S1).

For most simulations we use a 1 Mbp genome and n = 10 sample size. For the atlas (Table S1), for computational efficiency, we use a reduced genome (100 Kbp) and sample size (n = 5), which is sufficient to characterize the nucleotide divergence as shown in Fig. S7.

Recruitment

We randomly sample a cell from one location and obtain its SCG. Then, for another location, we sample fragments ($n = 10^4$) by repeatedly selecting a cell and a random piece from its genome (fragment, l = 1,000 bp). Global Ocean Sampling (GOS) Expedition samples included on average 1.5×10^5 fragments with 920 bp length (*7*). Recruitment is done by BLAST alignment at a specified identity (e.g., 99.9%).

For the OTU provinces mapping (Fig. 3), not all OTUs may be represented in the sample of SCGs (n = 100) and grid boxes are only assigned to an OTU if the fraction recruited corresponds to the fraction of one OTU. For example, if there are 5 OTUs, then the maximum recruitment fraction from a grid box has to be at least 1/5 for it to be assigned to one of these OTUs.

Supplementary Text

S1. Effect of population size on biogeographic patterns

Computational constraints prevent us from running the global model with a realistic number of individuals. For example, a population of prokaryotes in the surface ocean of 3.6×10^{28} cells (*25*) with a 1.3×10^{6} bp (*34*) genome stored at 4 bp per byte would require twelve billion yottabytes of memory. This is a common problem in agent-based modeling of microbes, which is typically addressed by simulating a smaller number of representative cells, referred to as super-individuals (*21*). This approach prohibits us from a direct simulation of absolute diversity. However, biogeographic patterns, as quantified here, are independent of the number of cells (population size). Here we review the theory and provide results from our model applied to a simple two-box system to support this.

Theory

Support is provided from neutral island model theory (*27, 39*), which is reviewed here in the context of our model. First, the number of mutations and difference of a cell from its ancestor (d', bp / cell) increases in a linear, clock-wise manner based on the mutation rate (k_u, bp / cell / generation) and time (t, generations) according to $d' = k_u\, t$ (see Fig. S3). The difference between two independently evolving cells that descend from the same ancestor (d) is twice that, $d = 2\, k_u\, t$. Therefore, the expected difference between two cells can be computed from the expected time to their common ancestor, the coalescence time (t_C, generations), which can be derived analytically for a number of scenarios. Here we consider the finite island model and assume infinite alleles, neutral mutations and equilibrium. Consistent with the application of our full model to a two-box system below, we use two subpopulations of size N (each) that interact by migration (k_m, cells / cells / generation). For two cells from the same island X (or Y) $t_{C,X} = 4\, N$ and $\pi_X = 2\, k_u\, t_{C,X} = 8\, k_u\, N$. For two cells from different islands $t_{C,XY} = 4\, N + 1 / (2\, k_m)$ and $\pi_{XY} = 2\, k_u\, t_{C,XY} = 8\, k_u\, N + k_u / k_m$. The difference attributed to dispersal limitation can be computed as $\delta_{XY} = \pi_{XY} - (\pi_X + \pi_Y) / 2 = k_u / k_m$, which is independent of the population size. We present these steady-state theoretical estimates for π_{XY}, π_X, π_Y and δ_{XY} along with our model results in Fig. S9.

Application of full model

To confirm our model is consistent with the theory, we apply it to a simplified laboratory-scale, two-box reactor system with a range of cell numbers, from 10^5 down to 10^2. This 3 order of magnitude range is still far below the ~23 order of magnitude downscaling for the full model (10^5 vs. 3.6×10^{28} cells (*25*)), but it is substantial and provides some insight into the effect of computational resolution on biogeographic patterns.

In these simulations, cells divide, die and mutate as in the global model and are transported (migrate) between the boxes at a specified rate ($k'_m = 1\times10^{-5}$ year^{-1}). The carrying capacity was adjusted to control the number of cells in the simulation.

We perform a number of simulations for each population size and present the averages (Fig. S9B). As expected, the total nucleotide difference (π_{XY}) increases for the simulations with higher cell numbers. This can be attributed to the higher local nucleotide diversity (π_X, π_Y) within each box (Fig. S9C). However, the nucleotide divergence (δ_{XY}), calculated as the difference between total nucleotide difference (π_{XY}) and local nucleotide diversity (π_X, π_Y) (see Methods) is relatively constant and independent of population size, confirming the theory laid out above.

For the higher population sizes, the total nucleotide difference (π_{XY}) and local nucleotide diversity (π_X, π_Y) in the full model are below the theoretical steady-state estimate. That is because the full model has not reached a steady-state in the 1,500 year simulation period. This is confirmed by comparing the full model to an analytical solution for local nucleotide diversity (π_X, π_Y) vs. time (*40*) (Fig. S9C).

Most importantly, the nucleotide divergence predicted by the full model is independent of the population size and consistent with the theoretical estimate. Even when the local nucleotide diversity is very low, at the lower population sizes, the model correctly predicts nucleotide divergence.

S2. Additional results and discussion for long-term simulation

For the simulation starting diverse without mutation, each cell can be assigned an OTU and mapping them reveals the development of provinces (Fig. S2). The provinces mix as evidenced by the presence of cells from different OTUs at the same location (e.g., Fig. S2, 100,000 years, Southern Ocean and Central Atlantic subpopulations). Evidence for mixing among provinces is also provided by a long-term individual history traversing multiple provinces (Fig. 1C) and spikes in nucleotide divergence or metagenomics fragments caused by vagrant cells (Figs. 2, S8A). Provinces are subject to continuous invasion by cells from neighbors, which periodically leads to takeovers or coalescence (e.g., Fig. S2, North and Central Pacific between times 500 and 100k years, see also Movie S1).

At the end of the 100k year simulation period the model includes 2 OTUs (Fig. S2) in the Southern Ocean and everywhere else. Typical time scales of water to cross basins are decades to centuries (*41, 42*), so it is expected that different regions in the ocean are not very well connected. The subtropical front, which is the southern extent of the subtropical gyres in the Southern Hemisphere, limits the mixing between the Southern Ocean and the subtropical basins in the Atlantic, Indian and Pacific Oceans (*43*).

S3. Metagenomics fragment recruitment at two locations

An alternative way to quantify biogeographic patterns (vs. nucleotide divergence, Fig. 2) is to perform metagenomics fragment recruitment, which shows an inverse pattern

to nucleotide divergence (Fig. S8A). At any one time, the genomes of the cells from the two locations align with a relatively constant identity, so they either meet or do not meet the taxonomic cutoff, which would suggest a more step-wise (vs. gradual) decrease in recruitment in between coalescence events. However, fragment recruitment can resolve the gradual divergence of the genome, because individual fragments are from a location on the genome with or without mutation, and the fraction of fragments from un-mutated regions decreases in a gradual manner. The recruitment fraction is much higher than for observations (e.g., (*8*)), which is a reflection of the low local nucleotide diversity within each of the provinces (in the model).

The fraction recruited at any one time depends on the taxonomic cutoff (Fig. S8B, corresponds to Fig. S8A at 1,400 years). The simulation starting diverse without mutation has a constant fraction of 1.0. For these two locations, after the first takeover, all fragments are from cells that descend from the same lineage (i.e., initial cell) as the SCG and are therefore 100% identical. This does not change with further coalescence events. At this time, the simulations starting diverse and uniform with mutation have converged and show the same pattern (Fig. S8A, 1,400 years). At a relatively high taxonomic resolution (99.9%), the diversity is less than 1.0, which is due to the independent evolution of these two provinces, as discussed in the main paper. However, as the taxonomic resolution is decreased the recruited fraction increases. Below 99.5%, the recruited fraction is close to 1.0. The rate of mutations is too low to develop distinct populations (i.e., that are less than 99.5% identical) in the time since the previous coalescence event.

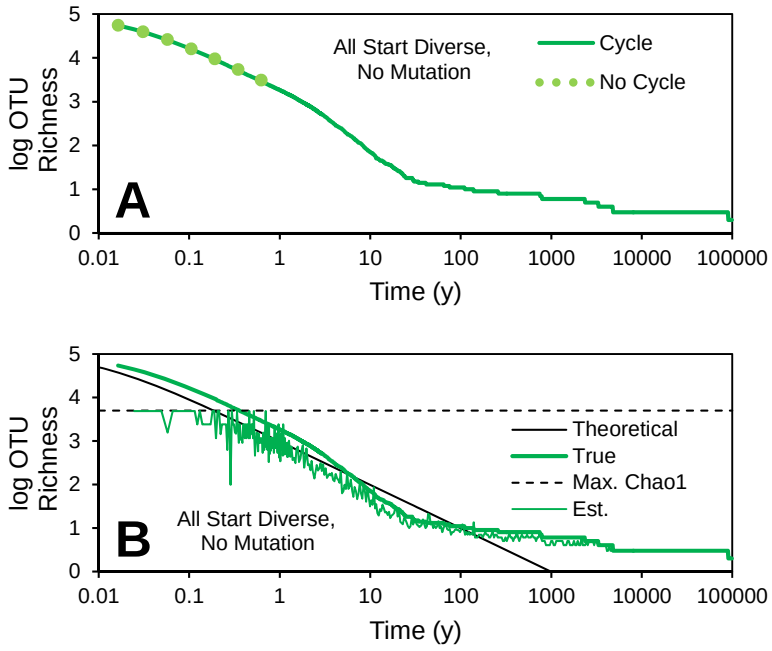

Fig. S1.

OTU richness over time (as in Fig. 1A). (**A**) Effect of history cycling. "Cycle" and "No Cycle" refer to cycling (see Methods). True richness based on all cells. (**B**) Comparison of two methods of quantifying diversity. "True": True richness based on all cells. "Max. Chao1": Maximum Chao1 richness estimate for a sample of $n = 10^2$ (i.e., $S_{obs} = 99$, $F_1 = 99$, $F_2 = 1$, see equation in Methods). "Est.": Estimated richness based on 99.9% whole-genome BLAST identity from a sample of $n = 10^2$ cells from across the globe using Chao1.

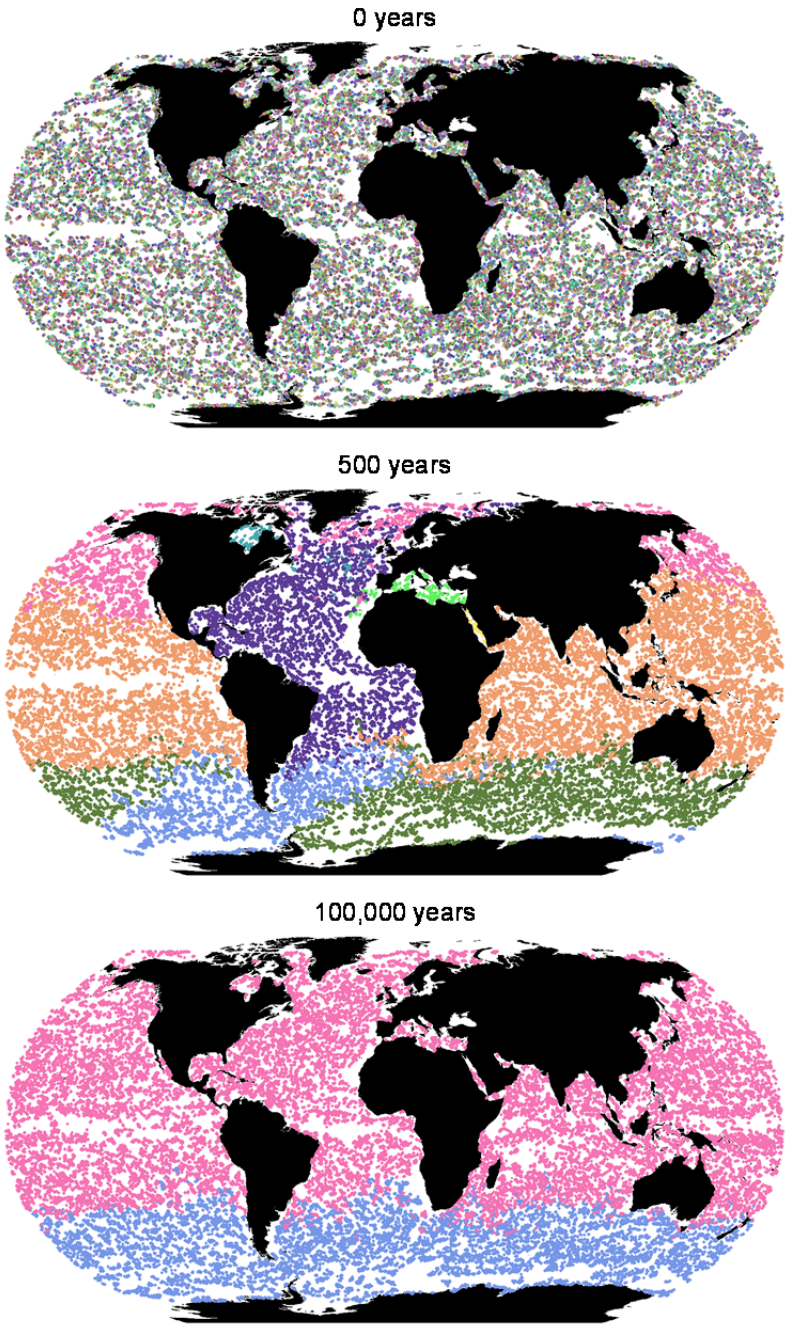

Fig. S2

Biogeographic patterns at various times. Simulation starting diverse without mutation. All model cells colored by OTU at various times. See also Movie S1.

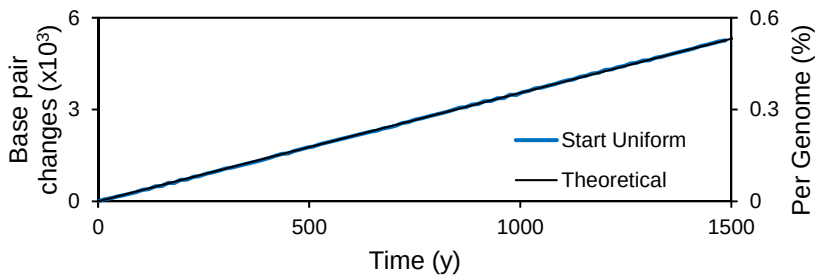

Fig. S3

Molecular clock. Accumulation of base pair changes over time. "Start Uniform": All initial cells have the same, completely random genome (same simulation as shown in Fig. 1B). Model values calculated by BLAST comparison of initial strain to $n = 10^2$ descendant cells sampled across the globe at various times. “Theoretical”: Based on neutral theory.

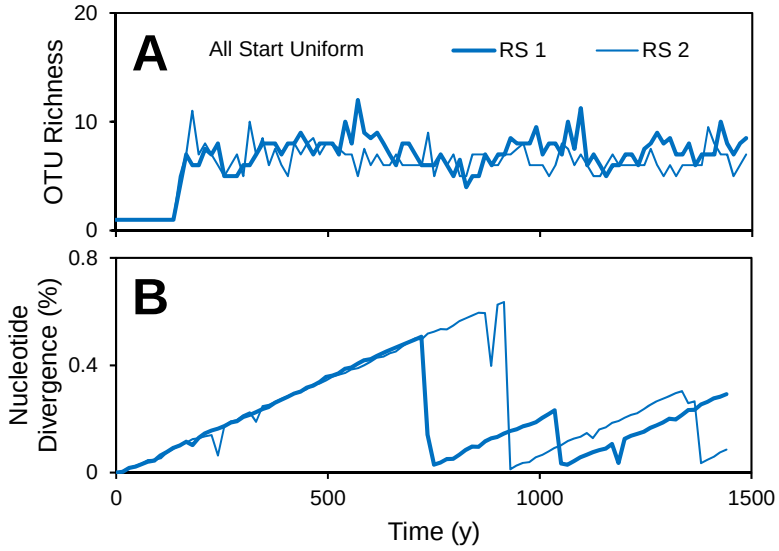

Fig. S4

Effect of stochasticity. Simulations presented in the paper use the same transport and sampling. This is evident in the simulations starting diverse and uniform with mutation in Fig. 1B (e.g., the spike in richness at 550 years) and Fig. 2 (coincidence of coalescence event at 1,050 years). This figure presents results from two simulations with different seed values for the random number generator (random seed, RS). (**A**) OTU richness over time (as in Fig. 1B). (**B**) Nucleotide divergence (biogeographic pattern) for two locations over time (as in Fig. 2A).

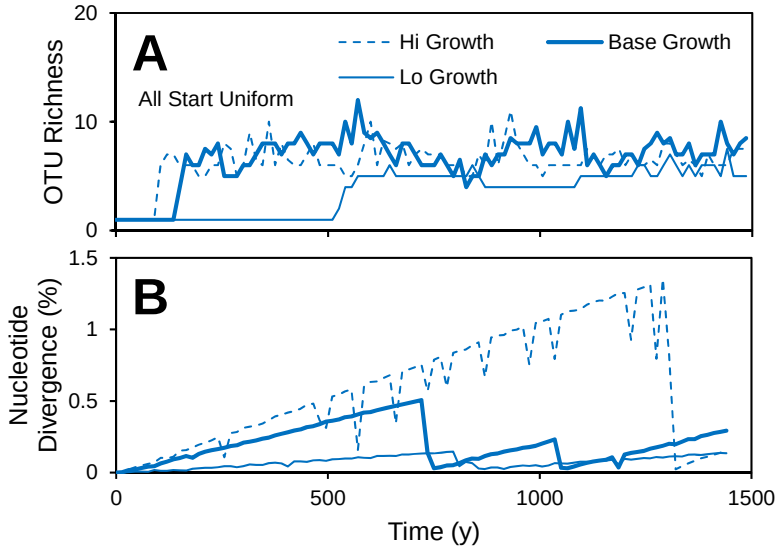

Fig. S5

Effect of growth rate. This figure presents results from three simulations with different growth rates, including the rate used in all other simulations ("Base", 0.14 day^{-1}), a low rate ("Lo", 0.05 day^{-1}) and a high rate ("Hi", 0.30 day^{-1}), based on ref. (*27*). (**A**) OTU richness over time (as in Fig. 1B). (**B**) Nucleotide divergence (biogeographic pattern) for two locations over time (as in Fig. 2A).

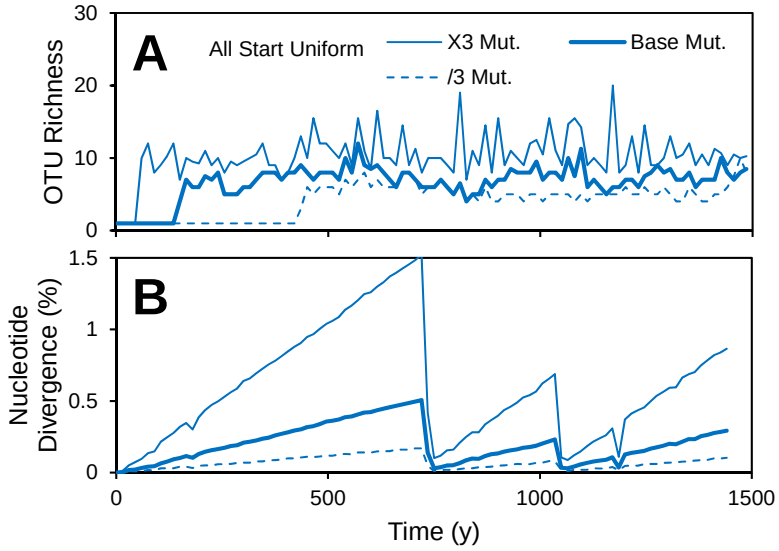

Fig. S6

Effect of mutation rate. This figure presents results from three simulations with different mutation rates, including the rate used in all other simulations ("Base", 3.5×10^{-8} bp/division), a low rate ("/3", 1.2×10^{-8} bp/division) and a high rate ("X3", 1.0×10^{-7} bp/division). (**A**) OTU richness over time (as in Fig. 1B). (**B**) Nucleotide divergence (biogeographic pattern) for two locations over time (as in Fig. 2A).

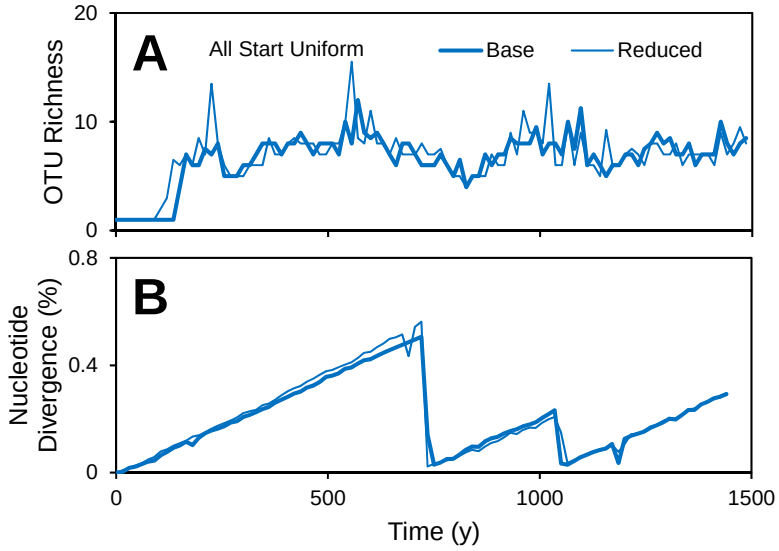

Fig. S7

Effect of reduced genome and sample size. This figure presents results from two simulations with different genome and sample sizes, including the values used in all other simulations ("Base", 1 Mbp, n = 10) and lower values used for the generation of the atlas (Table S1) ("Reduced", 100 Kbp, n = 5). (**A**) OTU richness over time (as in Fig. 1B). (**B**) Nucleotide divergence (biogeographic pattern) for two locations over time (as in Fig. 2A).

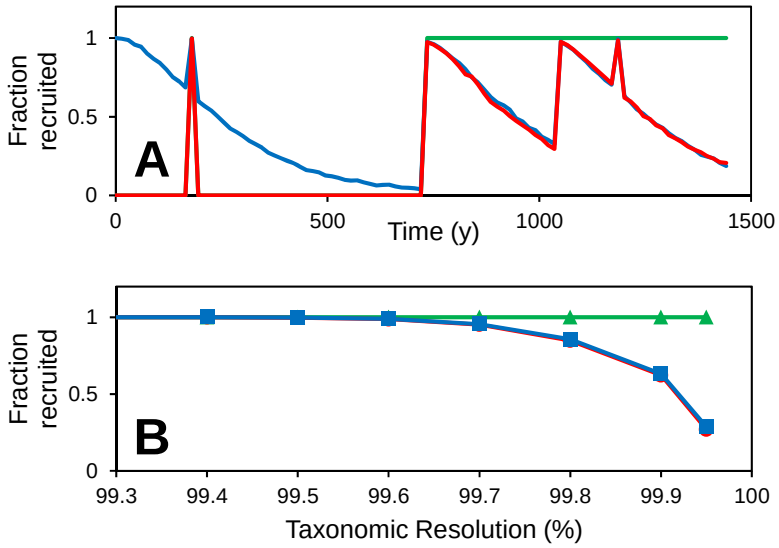

Fig. S8

Biogeographic pattern quantified by metagenomics fragment recruitment considering two locations. See caption Fig. 1 for description of simulations. (**A**) Recruitment over time. Fragments collected at GOA (n = 10,000, l = 1,000 bp) were recruited against an SCG collected at point HOT (n = 1). (**B**) Recruitment at various taxonomic resolutions. Same as panel A at 1,400 years.

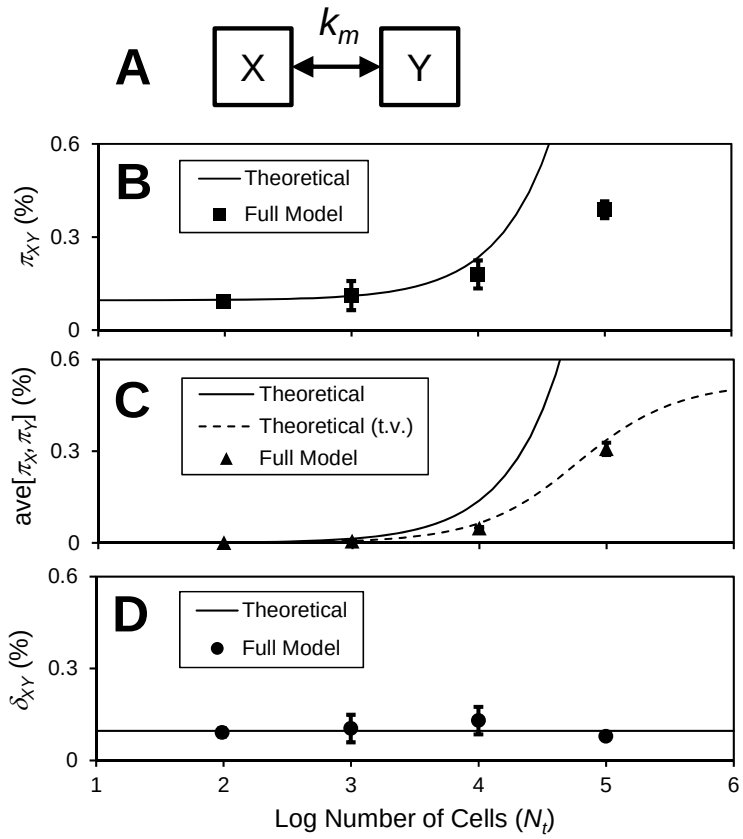

Fig. S9

Effect of population size ($N_t = 2\,N$). (**A**) The full model was used to simulate a two-box system using different number of cells. (**B**) Total nucleotide difference (π_{XY}). (**C**) Local nucleotide diversity (π_X, π_Y). (**D**) Nucleotide divergence (δ_{XY}), i.e., biogeographic pattern. SCGs from box X (n = 10) were aligned with SCGs from box Y (n = 10) at various times, as in the global model (Fig. 2). Symbols represent averages of n = 30 (for log N_t = 2-4) or n = 8 (for log N_t = 5) runs with different seed values for the random number generator, of averages over the 1,500 year simulation period. Error bars are ±2 standard error of the mean. "Theoretical" are steady-state theoretical values (see text). "Theoretical (t.v.)" are time variable theoretical values (*40*).

Table S1.

Atlas of neutral biogeography. DXYAVE and DXYMAX columns are mean and maximum of nucleotide divergence for two locations (δ_{XY}, as in Fig. 2). Start Uniform simulation, 1,500 years, with 100 Kbp genomes and $n = 5$ sample size (see Fig. S7 for the effect of reduced genome and sample size). Land-land and land-water pairs are omitted.

Movie S1

Map of all model cells colored by OTU. Simulation starting diverse without mutation. See also Fig. S2.