Shaw, J.O., Briggs, D.E.G., and Hull, P.M., 2020, Fossilization potential of marine assemblages and environments: Geology, v. 49, https://doi.org/10.1130/G47907.1

Fossilization potential of marine assemblages and environments

Jack O. Shaw[1*], Derek E. G. Briggs[1,2], and Pincelli M. Hull[1,2]
[1]Department of Earth and Planetary Sciences, Yale University, New Haven, Connecticut 06511, USA
[2]Peabody Museum of Natural History, Yale University, New Haven, Connecticut 06520, USA

Table of Contents

SUPPLEMENTARY INFORMATION

Data processing

We downloaded all metazoan occurrences from the OBIS on 21st January 2020 (Supplementary Data DR1) and retained only those species names recognized in the World Register of Marine Species (WoRMS) (Horton et al., 2020). We removed occurrences (i) lacking data on genus, latitude, longitude, depth, or duration of sampling, (ii) from non-marine settings, or (iii) with inadequate error control (e.g., citizen science data). Occurrences were grouped into assemblages if they were obtained within 0.1° latitude and 0.1° longitude of each other, in the same 10 m depth bin, during the same year, and from the same dataset. We retained only assemblages with representatives of at least three phyla, four classes, five orders, six families, and seven genera in order to avoid poorly sampled assemblages (Supplementary Data DR2). Alternative binning procedures were tested and did not affect interpretations (Fig. S1). Data processing and analyses were performed in R. All data are provided in the Data Repository (see pages 27-29).

Environmental assignments (shallow = continental shelf and above, deep = continental slope and below, coral reef, and seamount) were based on dataset descriptions and not inferences from faunal lists. We assigned substrate type (mud, sand, gravel, rock) to shallow assemblages (n = 7,545) based on data interpolated using K-nearest neighbor classification algorithms applied to the dbSEABED database (Jenkins, 2008) (Fig. S6). We did not assign substrates to other environments (e.g., deep water, seamounts, etc.) due to greater uncertainties introduced by the combined effects of lower data densities, interpolation method, and grid size.

We downloaded fossil occurrences from the PBDB on 17th January 2020 (Supplementary Data DR3). PBDB occurrences could not be assigned to substrate types due to a lack of sufficient PBDB information on substrate associations. PBDB occurrences were assigned to environments comparable to those in OBIS for the 25,732 genera with accompanying environmental data (Supplementary Data DR6). We used data on Konservat-Lagerstätten (i.e., exceptionally preserved fossil deposits listed in Muscente et al., 2017) which include 20,987 fossil occurrences of 4,398 genera only 208 of which have living representatives recorded in our subset of OBIS occurrences (n = 626,509) in order to test their impact on fossilization potential.

Pelagic and seamount assemblages

In the case of rarely preserved pelagic and seamount assemblages, low within-environment fossilization potentials reflect the rarity of these habitats in the rock record. Neither setting is recorded directly in the PBDB—we had to rely on the presence of 'seamount' and 'pelagic' in the data descriptions associated with PBDB occurrences. We identified seven geologic formations associated with seamounts in the PBDB, containing 483 occurrences and 213 genera. Seventy-eight percent of the OBIS genera exclusive to seamounts (253 of 5,635) are represented in the fossil record. Fifty-eight geologic formations in the PBDB are associated with dominantly pelagic taxa, containing 3,437 occurrences and 1,359 genera. Nine percent of the OBIS genera exclusive to pelagic assemblages (274 of 5,635) are represented in the fossil record. However, OBIS includes only 6 pelagic datasets (Fig. S8) and they include predominantly fish and/or zooplankton records and omit the abundant gelatinous members of the marine plankton. Thus the true taxon fossilization potential of pelagic environments is likely closer to their within-environment fossil potential of zero than that calculated here (Fig. 3, Fig. S3).

Taxon- and geography-specific analyses

Given the large focus on invertebrates in the PBDB, we performed analyses of fossilization potential excluding vertebrates from OBIS assemblages (Fig. S15). We found minimal changes (<3%) to mean taxon and within-environment fossilization potential at the global level and when data are parsed by substrate. There are also minimal changes (<5%) in shallow, deep, seamount, and pelagic mean fossilization potential values. Coral reefs have higher invertebrate-only taxon (from 44% to 54%) and within-environment (from 26% to 36%) mean fossilization potential values (Fig. S4).

In order determine how the spatial distribution of sampling and taxa—for both OBIS and PBDB data—impacted within-environment fossilization potential we analyzed a highly sampled area. Few areas are both well sampled and include assemblages covering a range of environment parameters, but we focused on an area with high PBDB and OBIS coverage, encompassing part of Europe (bounded by coordinates: 58°N, 15°W; 58°N, 22°E; 35°N, 15°W; 35°N, 22°E). Our sampling area contained areas of land and ocean in order to sample fossil and modern diversity. The area contained 8,945 OBIS assemblages, of which 5512 included information allowing within-environment fossilization potential to be calculated. All assemblages with environmental

information in the bounded area were shallow or deep water assemblages. Mean taxon fossilization potential was 34% and mean within-environment fossilization potential 30% (compared to 38% and 29%, respectively, for global data). Mean substrate taxon fossilization potential was lowest in rock and gravel-based assemblages (24% and 20%, respectively) and greatest in sand and mud-based assemblages (33% and 34%, respectively), conforming to the same trend of higher fossilization potential in finer grained substrates evident in the global data. Shallow water mean taxon fossilization potential was 33% and mean within-environment fossilization potential was 31% (compared to 34% and 32%, respectively, for global data). Deep water mean taxon fossilization potential was 27% and mean within-environment fossilization potential was 14% (compared to 34% and 15%, respectively, for global data). The greatest difference between the mean fossilization potential of global and European assemblages is the taxon fossilization potential of deep water assemblages, which is 7% lower in the European sample than the global sample (Fig. S9). This difference is unsurprising, given the large variation in deep water taxon fossilization potential. The consistencies between the European subset and the global data indicate that much of the trend in fossilization potential is robust to geographic sampling biases.

Sampling biases

OBIS assemblages and PBDB taxa are dominated by shallow-water representatives, and these OBIS assemblages generally show the highest fossilization potential values. To test the relationship between sampling and fossilization potential we performed a number of subsampling analyses. It is important to note that the fossilization potential metrics used here are inherently reliant on taxonomic diversity, such that subsampling would remove a key point of our study: the PBDB (and the fossil record) is biased against the fossilization of certain habitats.

We tested biases in sampling using the three best sampled groups in terms of number of OBIS assemblages and number of PBDB genera: shallow, coral reef, and deep water environments. We excluded seamount and pelagic communities because subsampling of such small sample sizes—in terms of both OBIS assemblages and PBDB taxa—would obscure any trends. The smallest number of assemblages by environment is ~170 (seamounts) and the smallest number of fossil taxa by environment is ~200 (seamounts).

We subsampled to the smallest number of OBIS assemblages by environment (1012, coral reefs) and subsampled to the smallest number of PBDB genera by environment (5927, deep water) to calculate mean fossilization potential values (Fig. S13). We repeated this process 1000 times to generate 1000 mean taxon and within-environment fossilization potential values for each environment. Subsampling data in this way showed that (1) mean taxon fossilization potential values generated using all assemblages (e.g., pictured in Fig. 3) are within the interquartile ranges of mean values of subsampled data (Fig. S13) when parsed by environment, and (2) mean within-environment fossilization potential values generated using all assemblages are greater than mean values of subsampled data due to artificially reduced fossil diversity in the latter, and (3) the larger the reduction in PBDB sample size, the greater the difference between all data mean within-environment fossilization potential values and subsampled data mean values. The within-environment fossilization potential calculation no longer serves its initial purpose of illustrating contrast between PBDB environments, as fossil diversity is maintained for the least sampled environment group but reduced for the other groups.

We also varied the number of PBDB taxa sampled by environment (as opposed to sampling to the smallest number of PBDB genera by environment, as described above) in order to consider how biases in fossil sampling impact within-environment fossilization potential values. This showed that, for a given level of fossil sampling, coral reef taxa are proportionally better represented in the PBDB compared to shallow and deep water fossil taxa (Fig. S14). Fossilization potential values at the PBDB sample size of 5927 (Fig. S14, x-axis) are the same as those depicted in within-environment boxplots in Fig. S13.

Taxon durations and fossilization potential

The duration of a taxon impacts the chances of it entering fossil record. A long-lived taxon is more likely to enter the fossil record than a short-ranging taxon as it has a greater chance of recovery. Taxon duration is thought to be determined primarily by geographic range (Jablonski and Hunt, 2006). Additionally proposed biotic and abiotic determinants of duration include habitat depth (Fortey, 1980; Jablonski and Bottjer, 1983), depth range, body size (Jablonski, 2008), life mode (Crampton et al., 2010), niche breadth (e.g., number of environments occupied: Kammer et al., 1997). Thus we expect taxon duration to vary between environments and to contribute to assemblage-specific fossilization potential.

We considered the influence of genus duration on fossilization potential in two ways: (1) comparing the durations of OBIS genera recorded in the PBDB between environments and substrates; and (2) including the assemblage-specific means of these durations in a predictive model of genus-level fossilization potential. The mean genus duration of an assemblage only included taxa recorded in the PBDB and excluded taxa without a fossil record.

To generate genus durations we utilized the Adaptive Beta method, which is designed to estimate the true stratigraphic range of a taxon based on the temporal distribution of its fossil occurrences (Wang et al., 2016). The code allowed only 150-200 occurrences per taxon to be included. For taxa with more than 150 occurrences in the PBDB we subsampled 150 occurrences (always including the oldest and youngest) 1000 times to generate 1000 estimated origination ages, and used the mean of these ages as the inferred origination age (Data Repository). For taxa with fewer than six occurrences we used the oldest occurrence as the origination age, and for taxa with fewer than 150 occurrences we applied the method without subsampling.

We assembled a subset all of the OBIS genera recorded in communities with environmental information (n = 9,669). We plotted these data in two forms: (1) boxplots of genus durations by occurrence (Fig. S10A; i.e., genera recorded in multiple OBIS environments were counted multiple times), and (2) boxplots of the assemblage-specific average genus durations (Fig. S10B).

We found that genus durations were similar across substrate types (Fig. S10A). Genus durations were also greatest in shallow water assemblages and lowest in seamount assemblages. When the genus durations of taxa within assemblages were averaged we found that shallow water assemblages have the highest mean values, whereas pelagic and seamount assemblages have the lowest values. Mean assemblage-specific genus durations were greater in finer substrates than in coarser substrates.

Using linear regressions, we found a significant—albeit weak—negative correlation between assemblage-specific mean genus duration and taxon fossilization potential (slope = -0.04, p-value << 0.05, adjusted R-squared = 0.04). We also found a significant, weak, positive correlation between mean genus duration and within-environment fossilization potential (slope = 0.02, p-value << 0.05, adjusted R-squared = 0.01). When parsed by environment or substrate we

found variable weakly positive and weakly negative correlations between genus duration and fossilization potential.

Inferred versus stratigraphic taxon durations

The difference between inferred taxon duration (calculated using the Adaptive Beta method) and stratigraphic taxon duration (calculated using the oldest known occurrence in the PBDB) is determined by the relative density of fossil occurrences—the greater the density of occurrences, the smaller the difference between inferred and stratigraphic taxon durations. Thus, any two environments might differ in average inferred taxon duration because either (1) true taxon duration, and thus inferred taxon duration, differs, or (2) true taxon duration is the same but preservation differs, and thus the estimates of inferred taxon duration differ. In the first case we would expect the sites with the best-preserved fossil assemblages to show the largest differences between inferred and stratigraphic durations. In the second case we would expect the best-preserved sites to show the smallest differences between inferred and stratigraphic durations.

We find that the difference between inferred genus duration and stratigraphic duration are greatest in shallow environments, followed by coral reefs and deep water environments, seamounts, and pelagic environments (Fig. S16). In other words, differences are greatest in the best-preserved environments, suggesting that we are identifying true differences in taxon durations between environments, rather than differences generated by sampling artifacts. Therefore we used inferred taxon duration as a proxy for true taxon duration in our analyses.

Predictive modeling

We used conditional inference classification trees (CTree) built in the R package *partykit* (Hothorn and Zeileis, 2015) to assess the relative importance of environmental differences in preservation potential (variables: environment, water depth, substrate, realm, alpha diversity), taxon longevity (variables: average genus duration), and sampling biases (variables: latitude, longitude, alpha diversity) on the fossilization potential of marine assemblages (Data Repository). CTree is a type of Classification and Regression Tree (CART) analysis that iteratively tests subsets of data until the null hypothesis of interdependence among variables cannot be rejected. Unlike some other classification tree models, it accounts for covariation among predictor variables for unbiased selection and identification of variable importance.

Additionally, the model was chosen because it handles overfitting, non-linear and non-parametric data, and missing information. Models were built using only communities with environmental data (i.e., shallow, coral reef, deep water, seamount, and pelagic communities; n = 9,738). We ran CTree analysis 1000 times for both types of fossilization potential. Relative variable importance scores, R-squared (R2), and root mean square error (RMSE) values were then averaged across the 1000 runs. Each run utilized a random sample of 70% of all assemblages. Variable significance was assessed at an alpha of 0.05 using Bonferroni adjusted Monte-Carlo p-values (10 permutations per analysis).

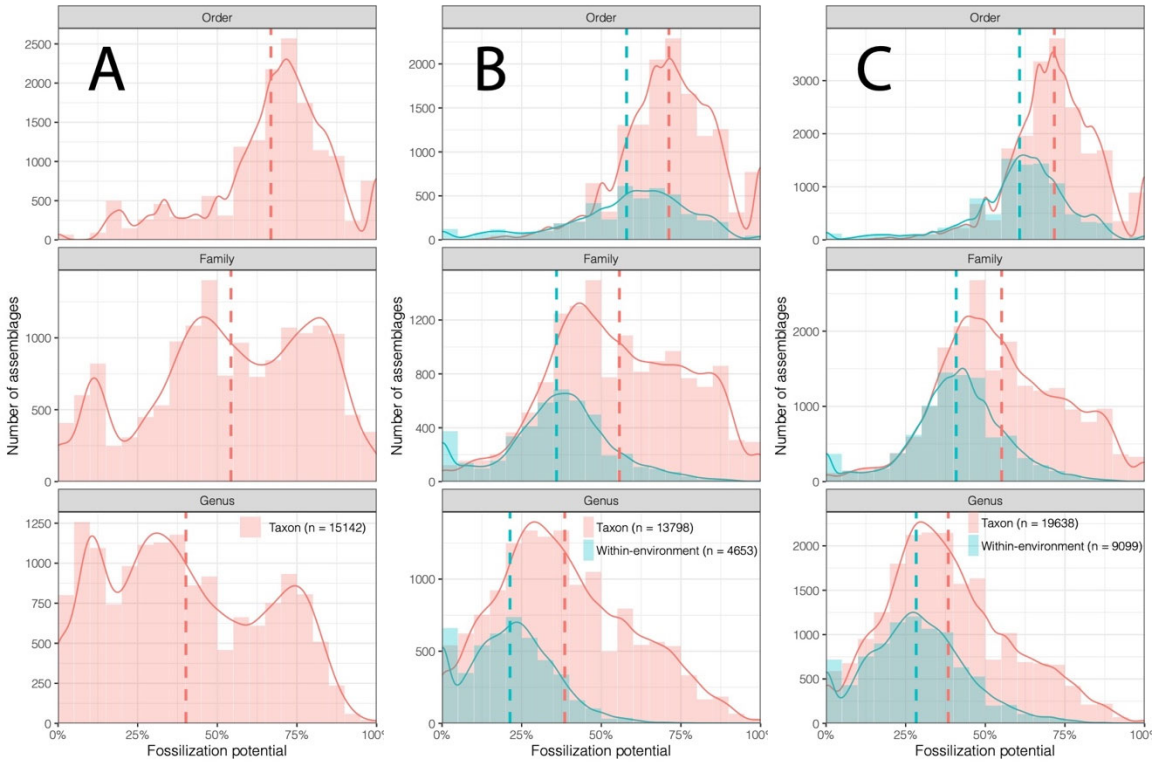
The dominance of spatial variables (longitude and latitude) in predictive models of fossilization potential is indicative of sampling bias. This is most obvious in analyses of pelagic community fossilization potential where different datasets focus on different taxa with very different fossilization potentials (Fig. S8). However, this sampling bias does not necessarily erroneously skew fossilization potential calculations. Most datasets are inherently spatially and environmentally restricted (i.e., one research group will focus on a small spatial patch, and not sample the globe), and these attributes covary (e.g., the presence of coral reefs around Australia).

Impact of Konservat-Lagerstätten on inferred fossilization potential of taxa

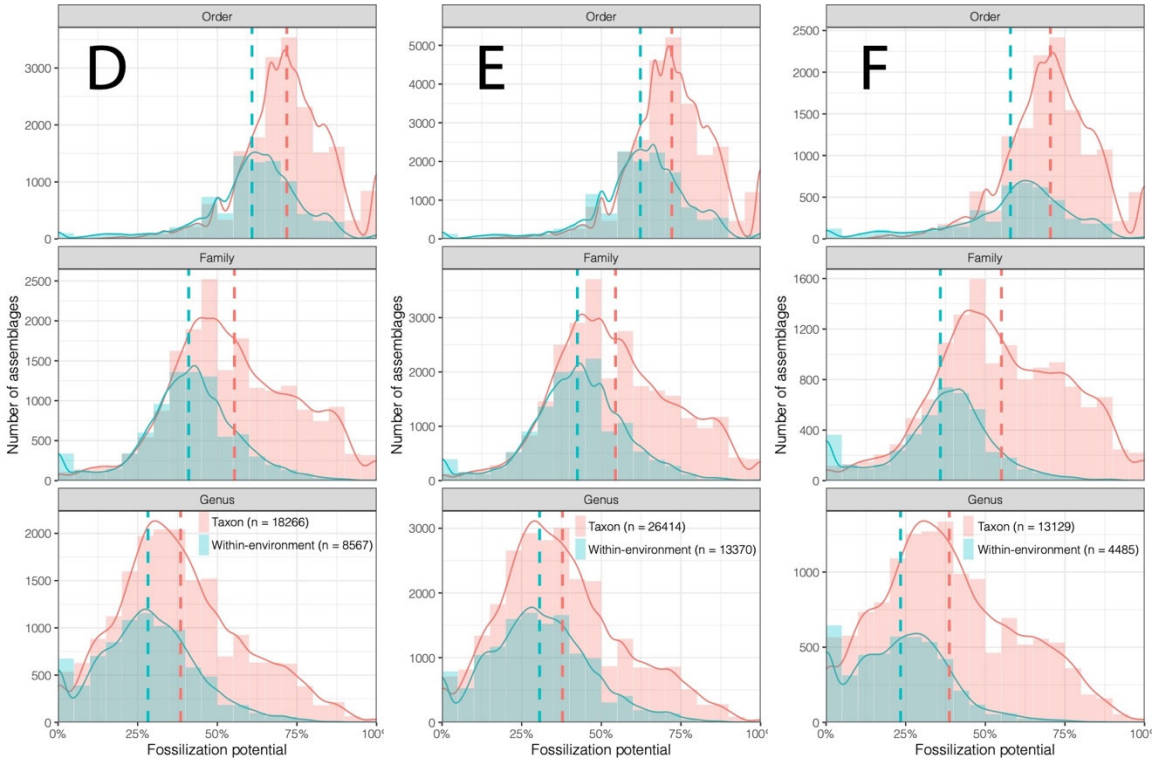
Taxa in Konservat-Lagerstätten add significant data to diversity and disparity estimates and our understanding of macroevolution particularly in the early Paleozoic (Briggs et al., 1994; Hou et al., 2017). Thus Konservat-Lagerstätten might be expected to have a dramatic effect on inferences of fossilization potential by providing rare glimpses of ancient assemblages. However, just 2.2% of the 58,870 genera in the PBDB are known only from Konservat-Lagerstätten. Furthermore, exceptionally preserved assemblages are much rarer in the Cenozoic (7 Konservat-Lagerstätten), when 75% (n=1,921) of OBIS genera with PBDB occurrences are found first, than in the Paleozoic (142 Konservat-Lagerstätten). Only 6 genera (0.05%, n=9,806) (7 families and 3 orders) in OBIS are recorded only in Konservat-Lagerstätten deposits and not in other PBDB deposits. When they are omitted from calculations of fossilization potential (i.e., limiting the study to ‘typical’ preservation conditions), there is no difference in our findings (Fig. S1). Thus our calculations of fossilization potential are robust to the inclusion (or exclusion) of Konservat-Lagerstätten from the PBDB – the shelly fossil record provides the primary archive of potential

assemblage composition. Particularly in the Cenozoic, there are few Konservat-Lagerstätten to fill known gaps in these biased archives.

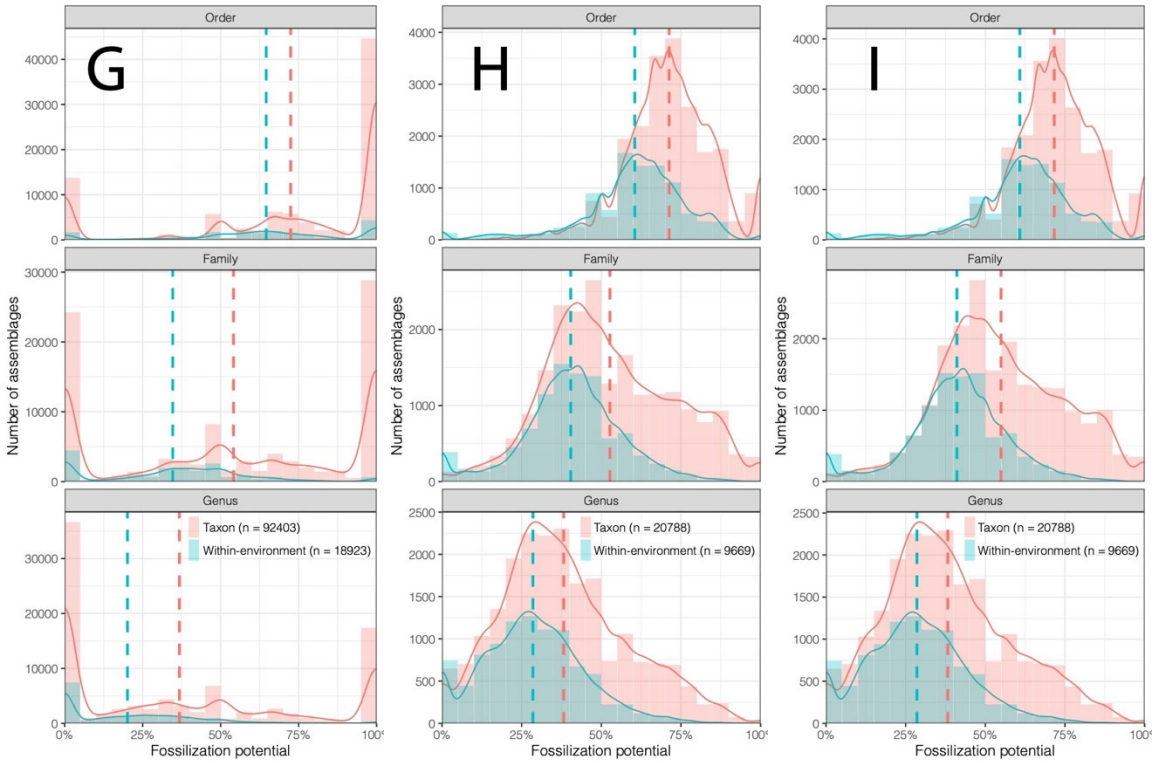
SUPPLEMENTARY FIGURES


A
B
C
Order
Family
Genus
Number of assemblages
Fossilization potential
Taxon (n = 15142)
Taxon (n = 13798)
Within-environment (n = 4653)
Taxon (n = 19638)
Within-environment (n = 9099)

234


D
E
F
Order
Family
Genus
Number of assemblages
Fossilization potential
Taxon (n = 18266)
Within-environment (n = 8567)
Taxon (n = 26414)
Within-environment (n = 13370)
Taxon (n = 13129)
Within-environment (n = 4485)

235

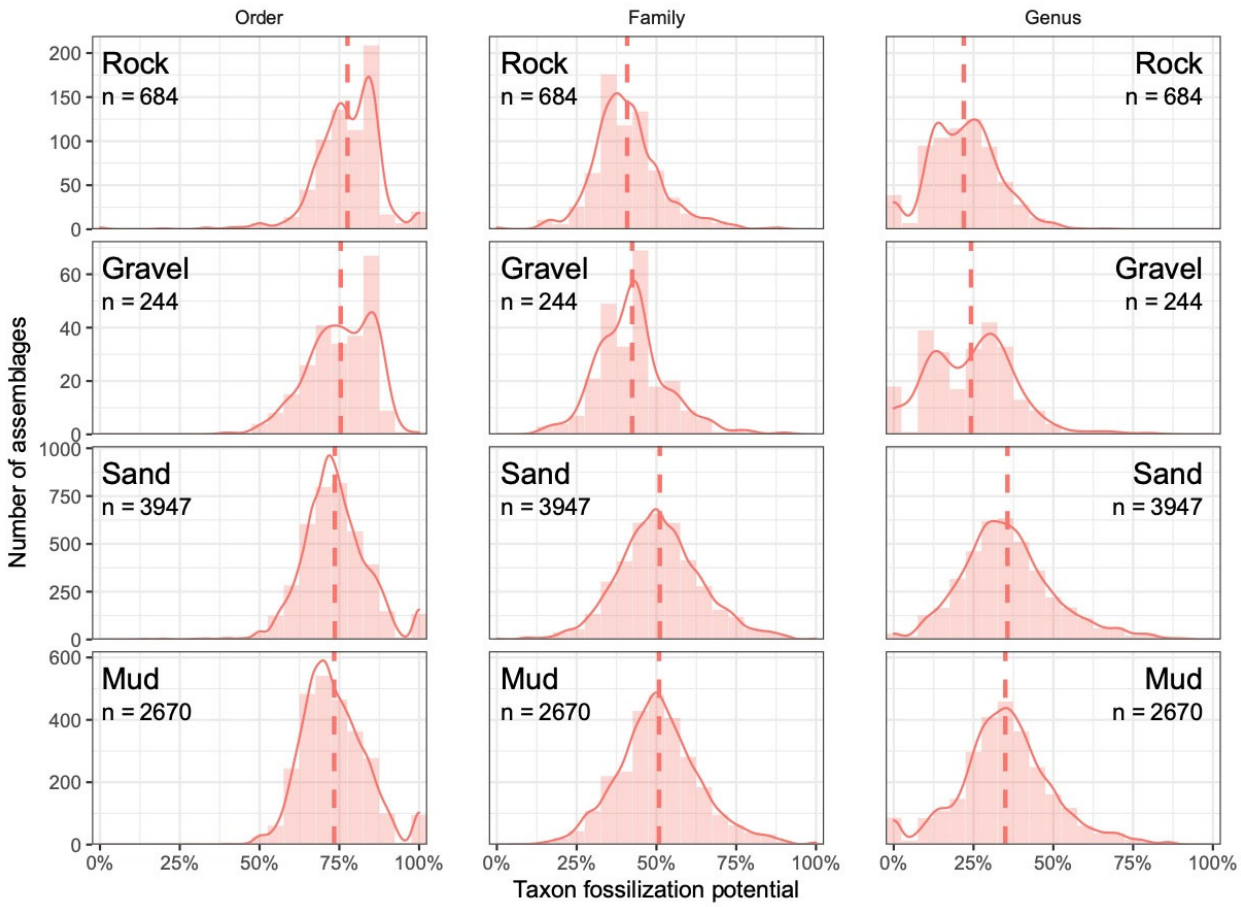


236

237 **Supplementary Figure S1:** Comparisons of taxon and within-environment fossilization
238 potential distributions calculated using alternative binning procedures. Unless otherwise noted
239 distributions include the default criteria used in the main text: all PBDB data and OBIS
240 assemblages grouped by 10 m depth bins, 0.1° latitude-longitude bins, sampling year, and a
241 minimum diversity of three phyla, four classes, five orders, six families, and seven genera.
242 Assemblages grouped with (A) 20 m depth bins and 1° latitude-longitude bins, but not by dataset
243 ID); (B) year excluded; (C) 20 m depth bins; (D) 100 m depth bins; (E) 0.01° latitude-longitude
244 bins; (F) 1° latitude-longitude bins; (G) no minimum diversity criteria; (H) Lagerstätten
245 excluded; (I) assemblages grouped using the default criteria (Genus panel: Fig. 1, main text).

246

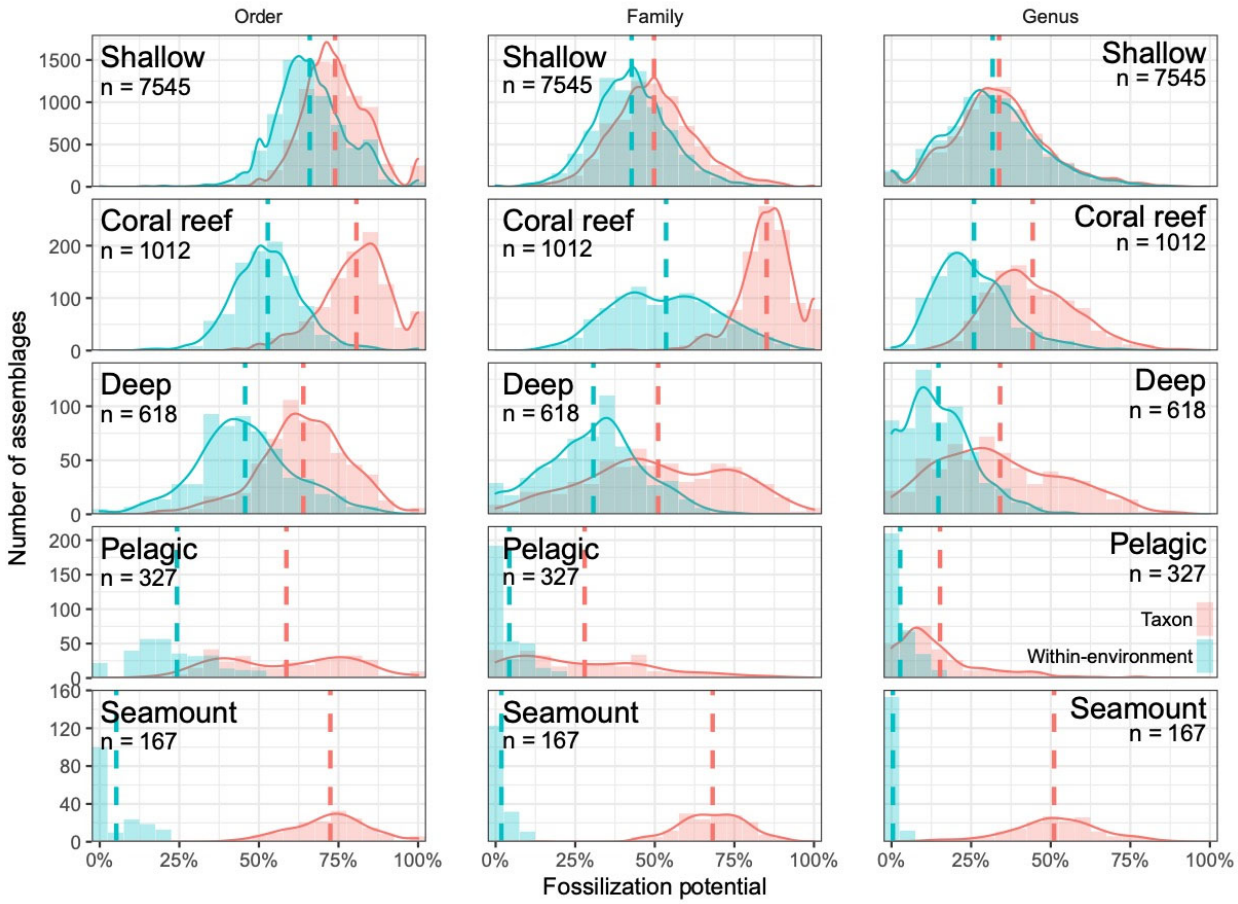
247



248

249 **Supplementary Figure S2:** Taxon fossilization potential distributions for shallow-water
250 assemblages with substrate information (n = 7,545). Taxon fossilization potential calculated at
251 ordinal, family, and generic ranks. Histogram bin widths = 5%, dashed lines indicate mean.

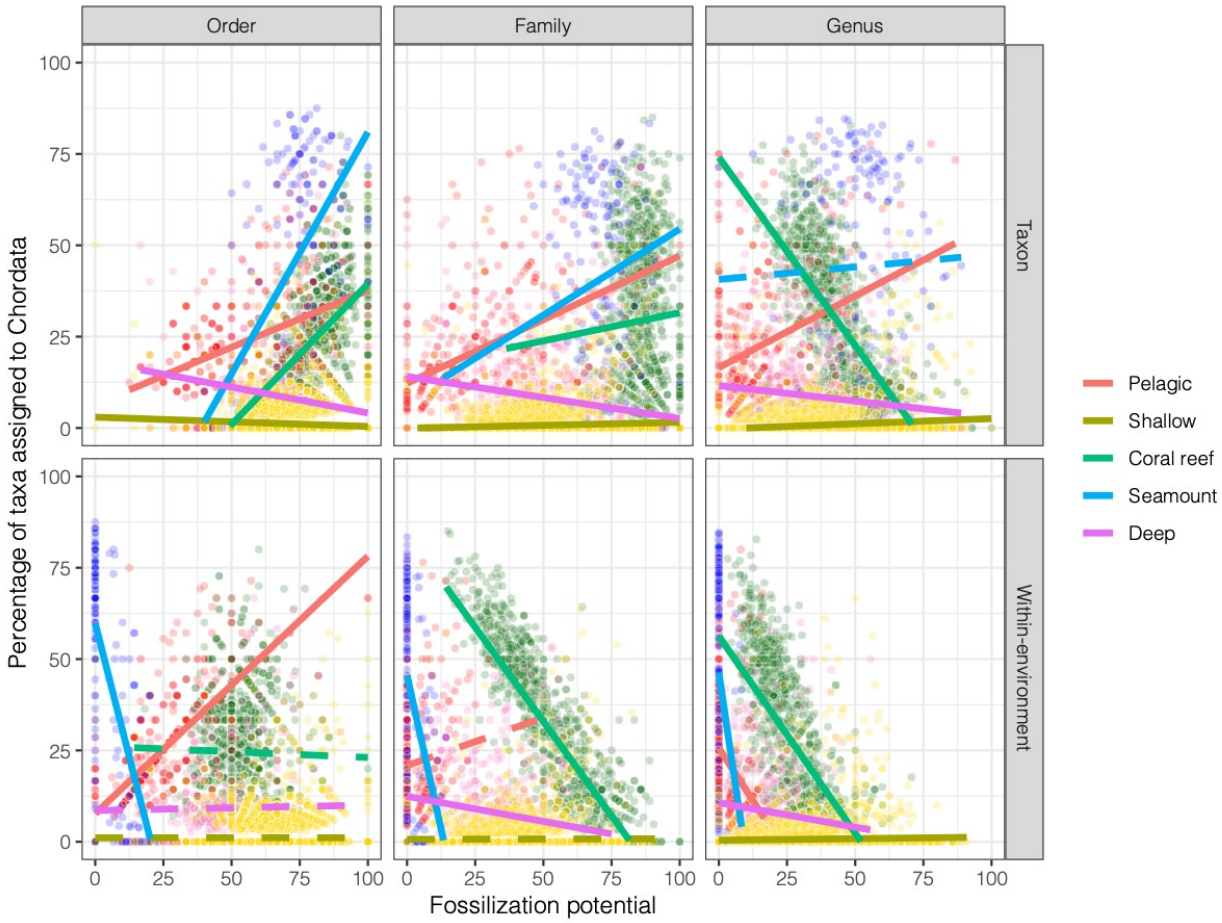
252



253

254 **Supplementary Figure S3:** Taxon (red) and within-environment (blue) fossilization potential
255 distributions of assemblages with environmental information (n = 9,669). Fossilization potential
256 calculated at order, family, and genus ranks. Histogram bin widths = 5%, dashed lines indicate
257 mean.

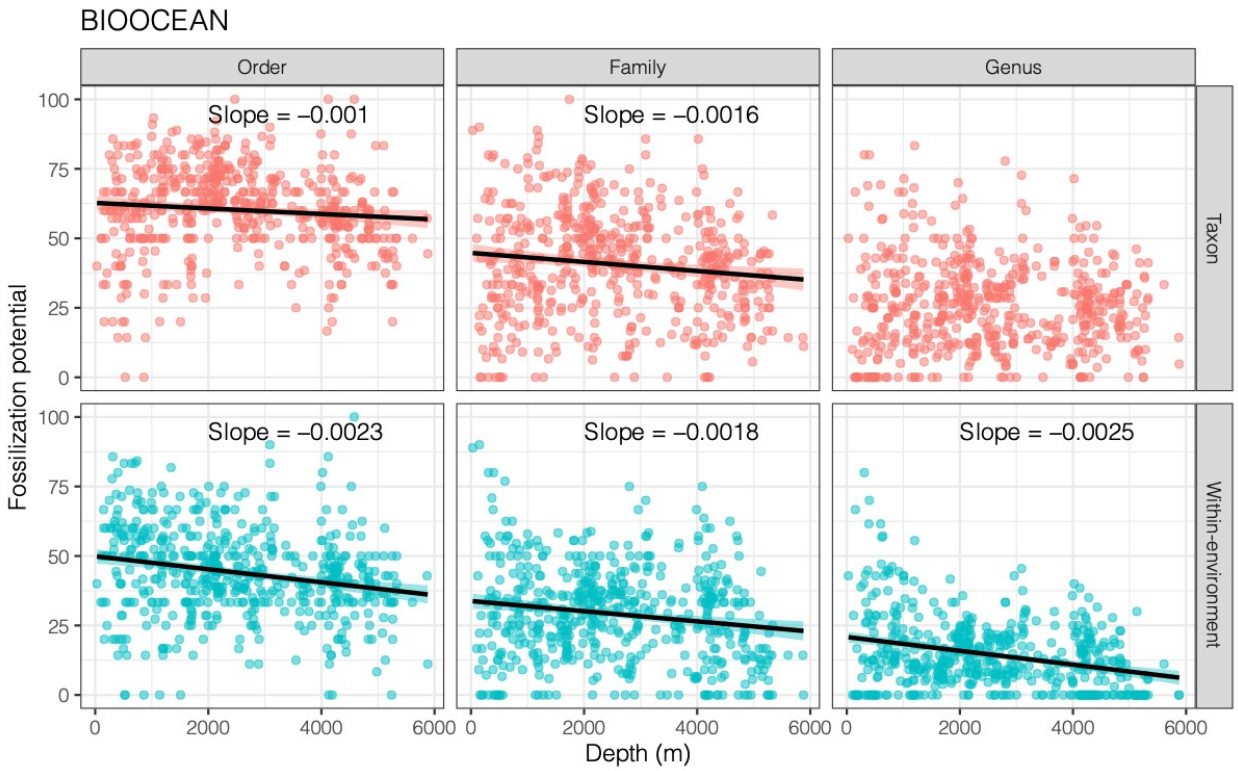
258



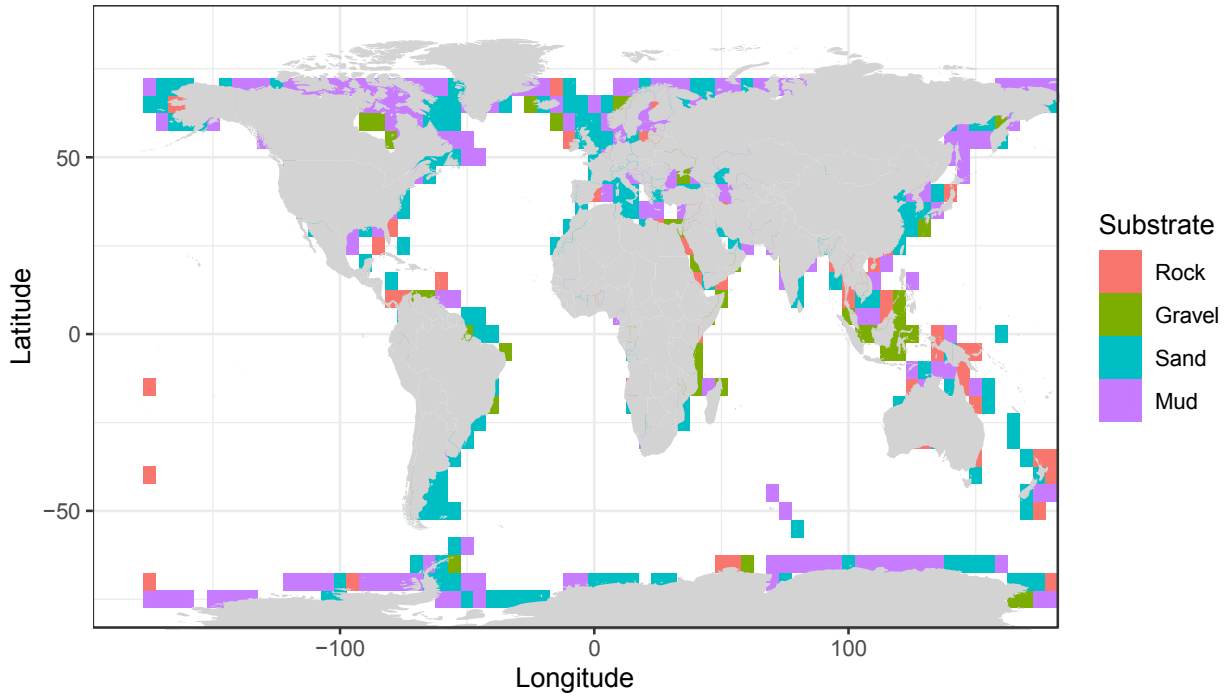
259

260 **Supplementary Figure S4:** Percentage of taxa (orders, families, or genera) in assemblages
261 belonging to the phylum Chordata, plotted against taxon and within-environment fossilization
262 potential. Datapoints indicate assemblages (n = 9,669) and are colored by environment. Solid
263 lines indicate significant linear regressions (p-value < 0.05), dashed lines indicate insignificant
264 regressions.

265



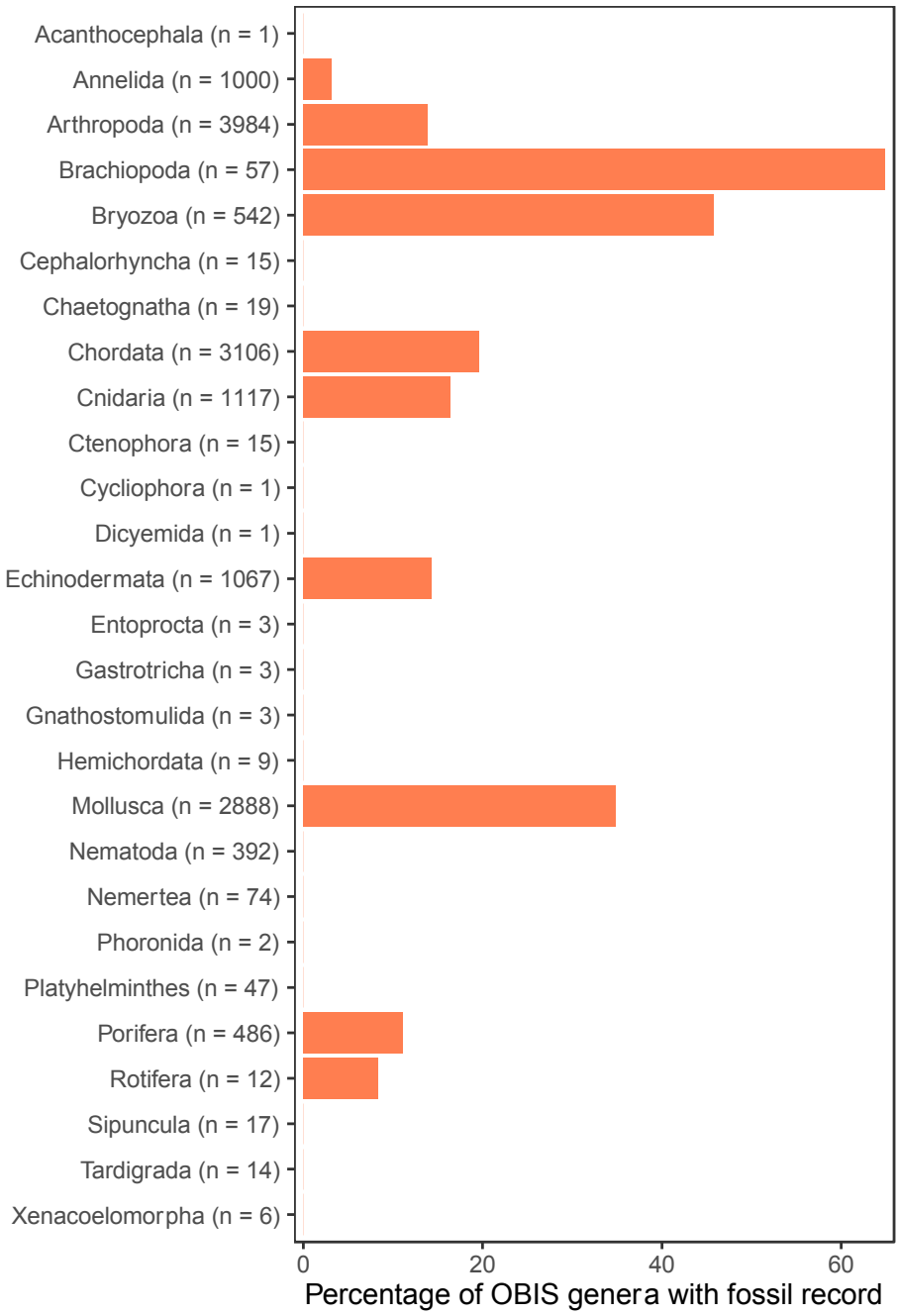
Supplementary Figure S5: Depth versus taxon (top) and within-environment (bottom) fossilization potential for data from the BIOCEAN dataset (n = 549; detailed in main text). Line shows statistically significant linear regression with slope values indicated.



271

272 **Supplementary Figure S6:** Map of shallow water (<1000m) substrate type interpolated using
273 dbSeabed data and KNN algorithm.

274

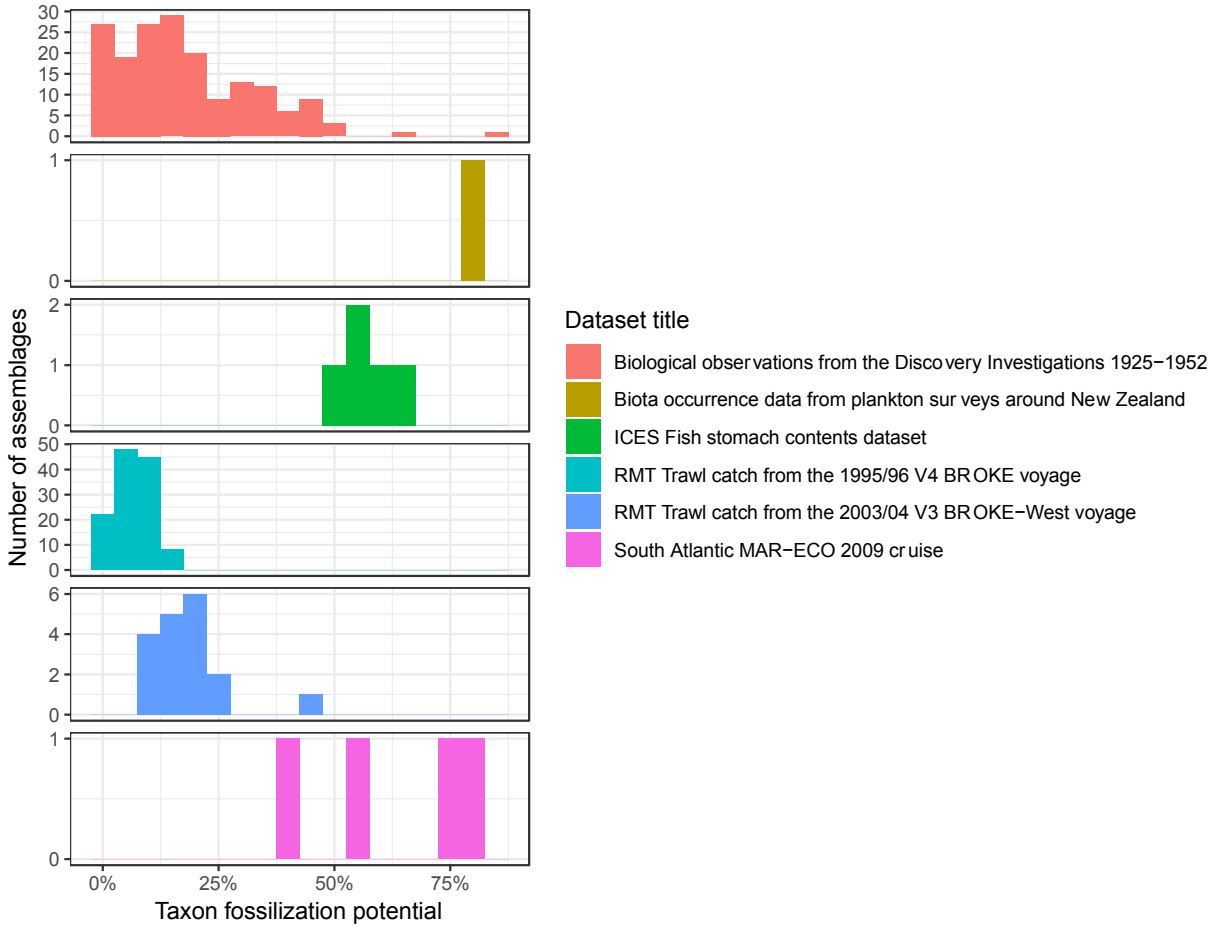


275

276 **Supplementary Figure S7:** Bar chart displaying the percentages of OBIS genera with records in
277 the Paleobiology Database (PBDB). The total numbers of OBIS genera are indicated in
278 parentheses.

279

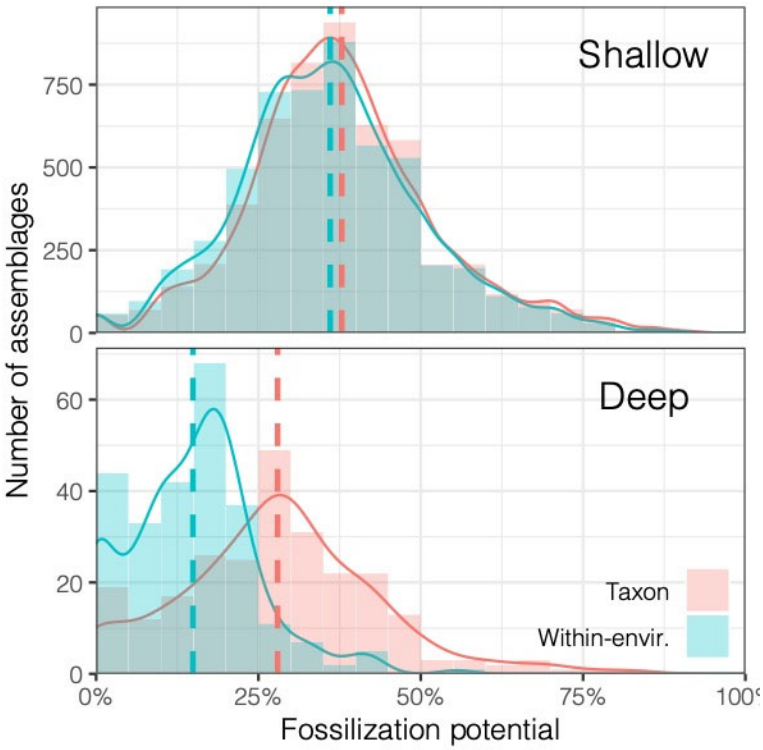
280



281

282 **Supplementary Figure S8:** Comparison of taxon fossilization potential distributions by dataset
283 ID (n = 6) for all pelagic communities (n = 327).

284

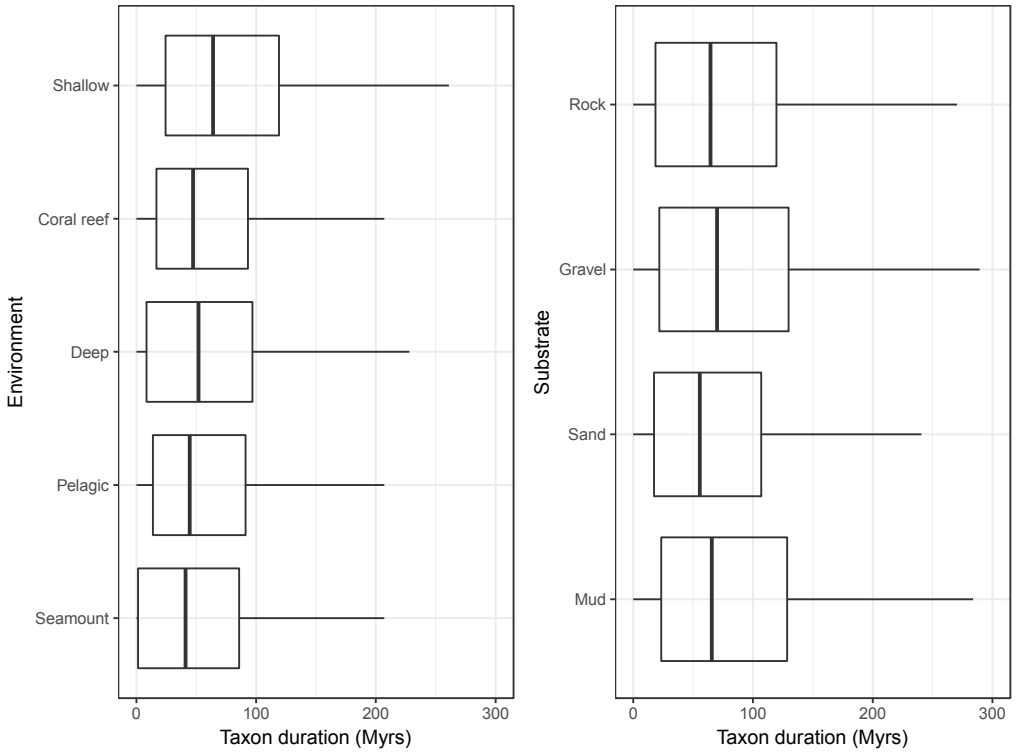


285

286 **Supplementary Figure S9:** Taxon and within-environment fossilization potential distributions
287 for assemblages within a bounded area in Europe (58°N, 15°W; 58°N, 22°E; 35°N, 15°W; 35°N,
288 22°E). Histogram bin widths = 5%; dashed lines indicate mean.

289

290



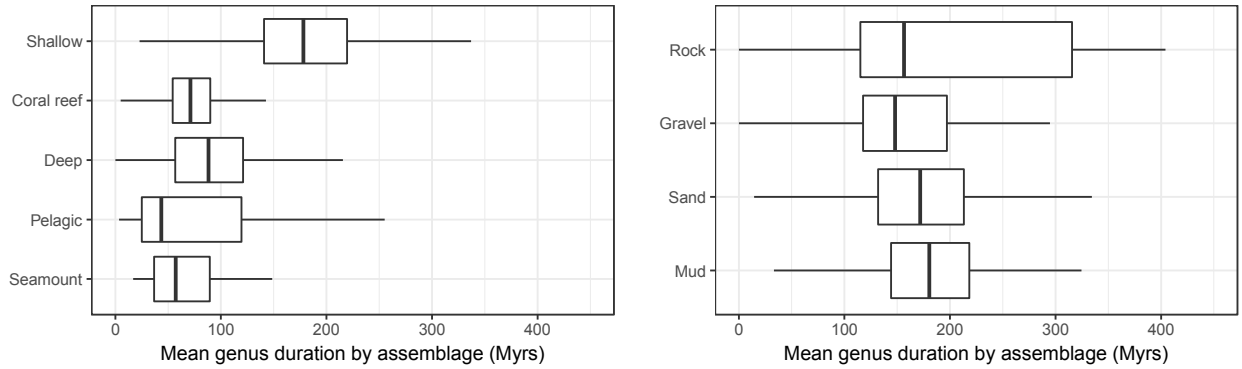
291

	Lower whisker	Lower hinge	Median	Upper hinge	Upper whisker
Shallow	0.01	24.36	64.06	119.12	261.10
Coral reef	0.01	16.60	47.31	93.64	207.05
Deep	0.01	8.47	51.90	96.89	228.06
Pelagic	0.01	13.79	44.55	91.23	207.05
Seamount	0.01	1.36	41.04	85.90	207.05
Rock	0.01	18.45	64.56	120.03	270.48
Gravel	0.01	21.74	70.17	129.77	289.32
Sand	0.01	17.32	55.63	107.12	240.76
Mud	0.01	23.54	65.65	128.57	283.93

292

293 **Supplementary Figure S10A:** Boxplots (top) and boxplot statistics (bottom) of genus durations
294 (calculation method outlined in supplementary information section "Taxon durations and
295 fossilization potential") for OBIS taxa with corresponding environment and substrate
296 information. Taxa without fossil evidence are excluded. Outliers removed for clarity.

297



298

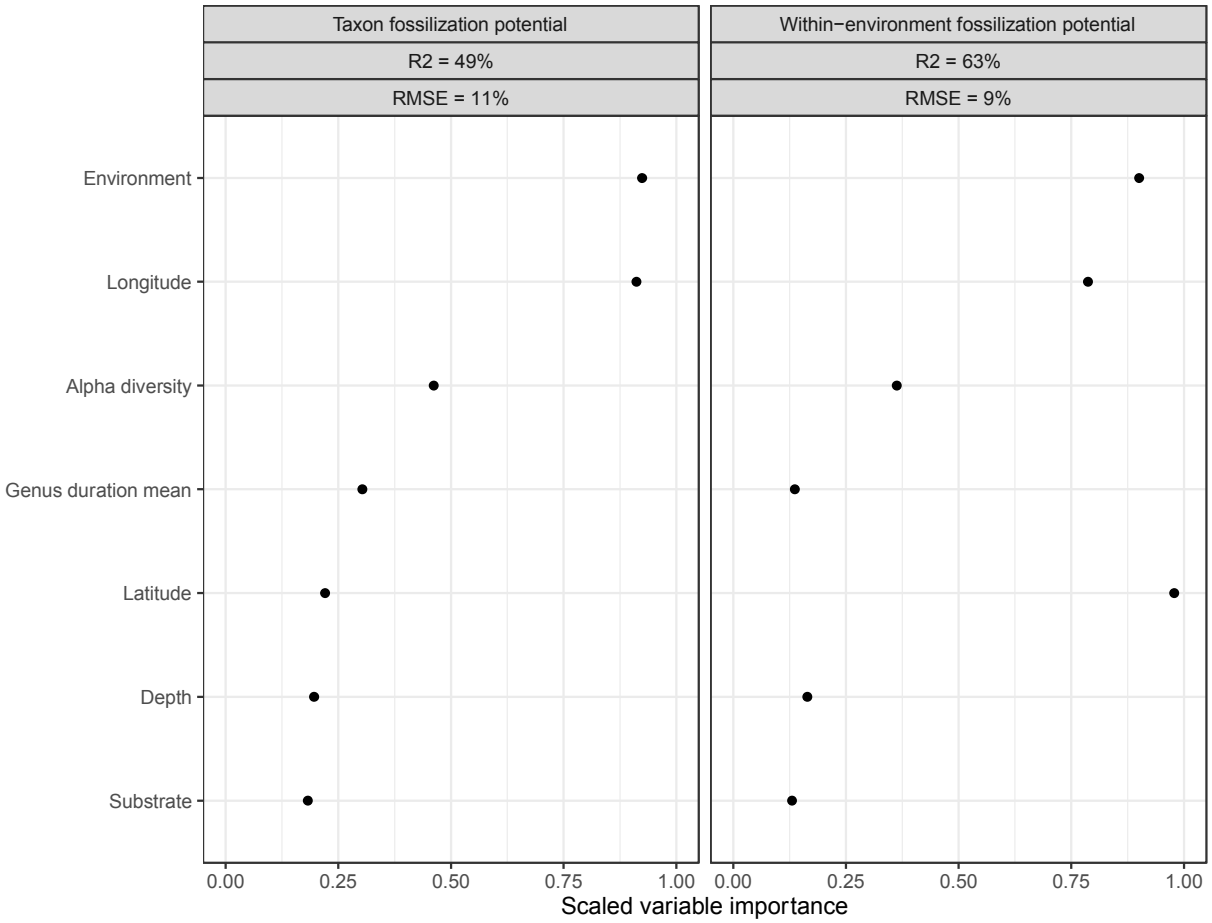
299

	Lower whisker	Lower hinge	Median	Upper hinge	Upper whisker
Shallow	22.95	140.81	178.16	219.43	337.13
Coral reef	5.08	54.30	71.19	89.94	142.76
Deep	0.01	56.60	88.24	121.07	215.55
Pelagic	3.49	25.02	43.52	120.13	261.92
Seamount	16.78	36.74	57.06	89.51	148.55
Rock	0.01	115.03	156.48	315.54	404.18
Gravel	0.01	117.52	147.91	197.38	294.72
Sand	14.39	132.00	171.75	213.15	334.38
Mud	33.15	144.18	180.40	218.34	324.65

300

301 **Supplementary Figure S10B** Boxplots (top) and boxplot statistics (bottom) of mean genus
302 duration values for assemblages with environment (top left) and substrate (top right) information.
303 Calculations of assemblage mean values only include the durations of taxa with fossil
304 representatives. Outliers removed for clarity.

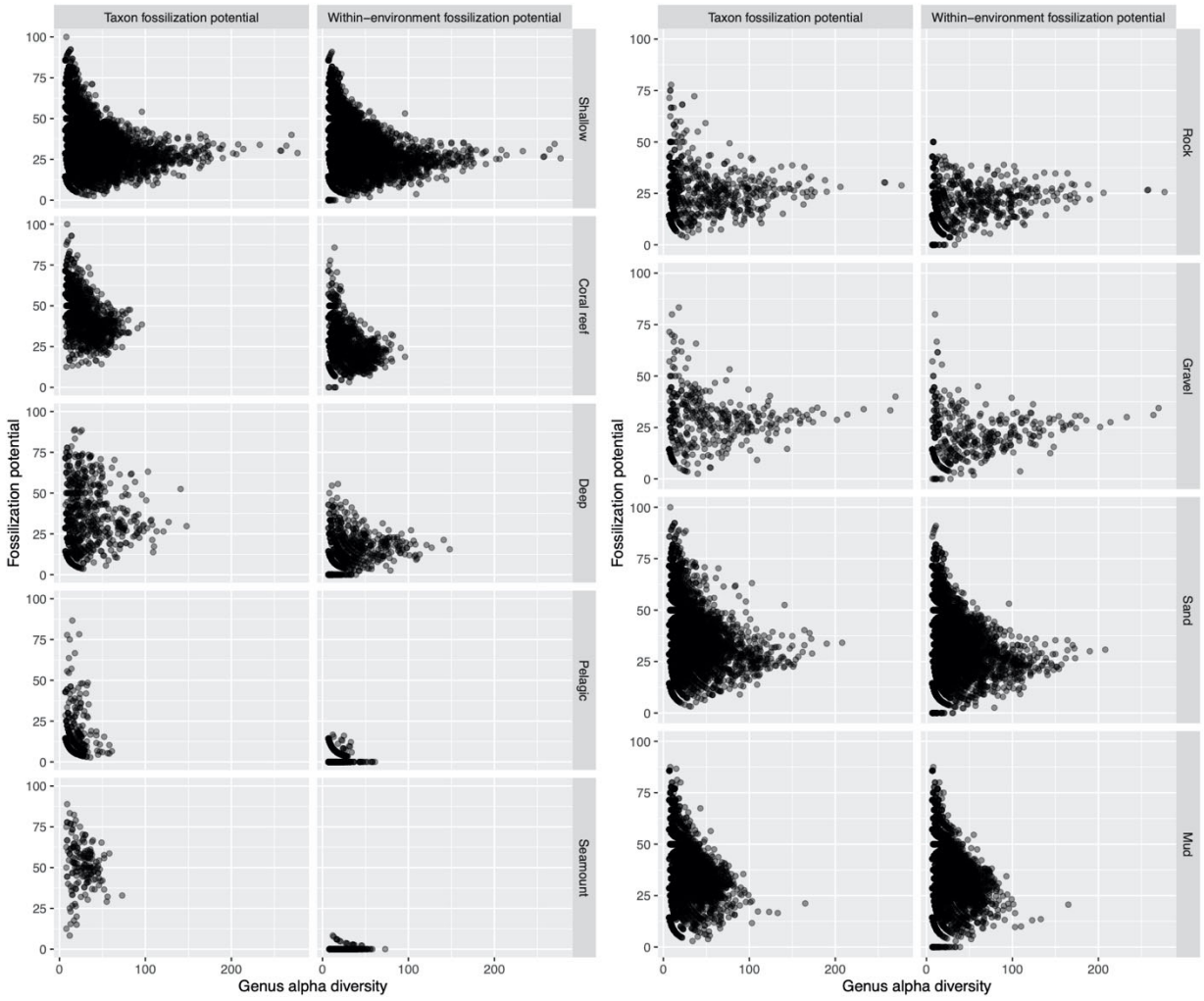
305



306

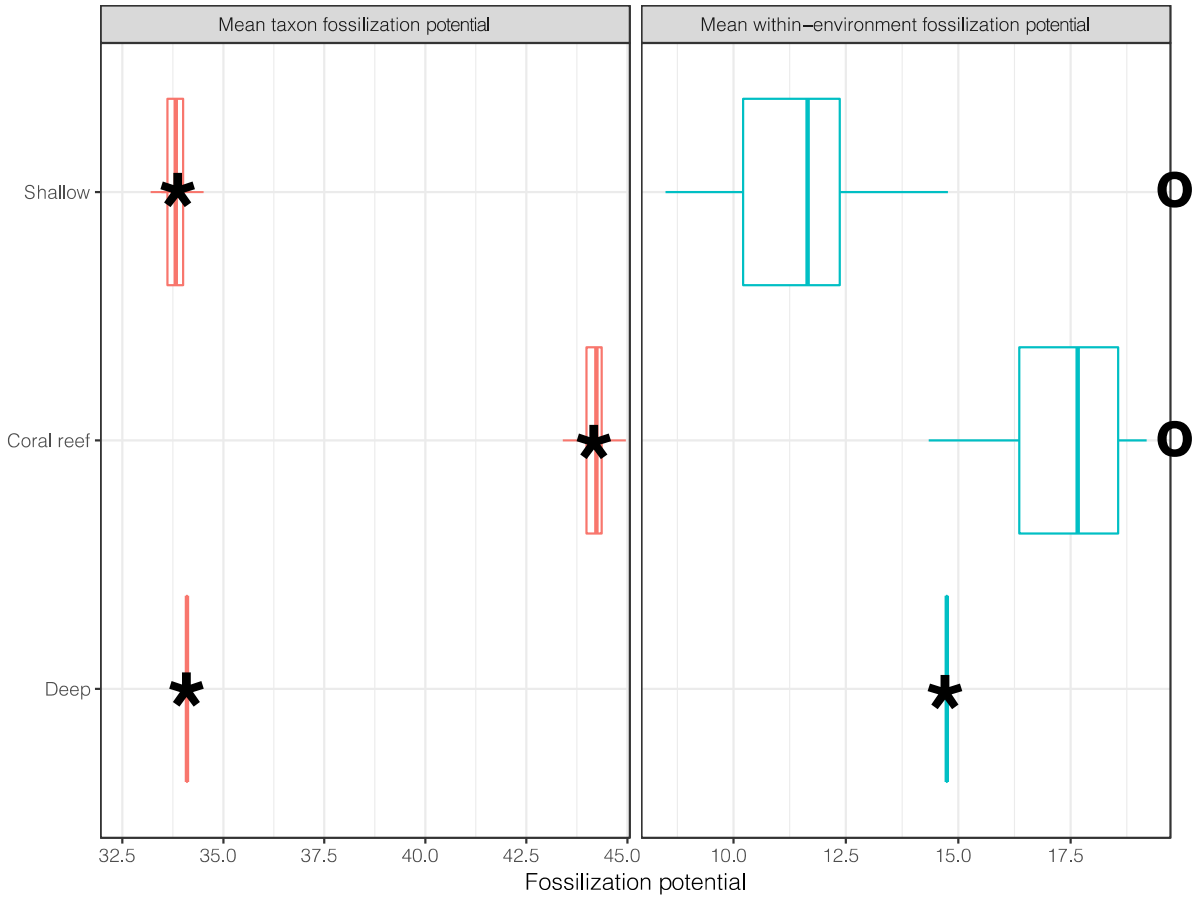
307 **Supplementary Figure S11:** Scaled variable importance scores for conditional inference
308 regression tree models predicting genus-level taxon and within-environment fossilization
309 potential. Importance rankings are based on the reduction in square error produced when the
310 variable is added. R-squared (R2) and root mean square error (RMSE) shown for each model.

311



312

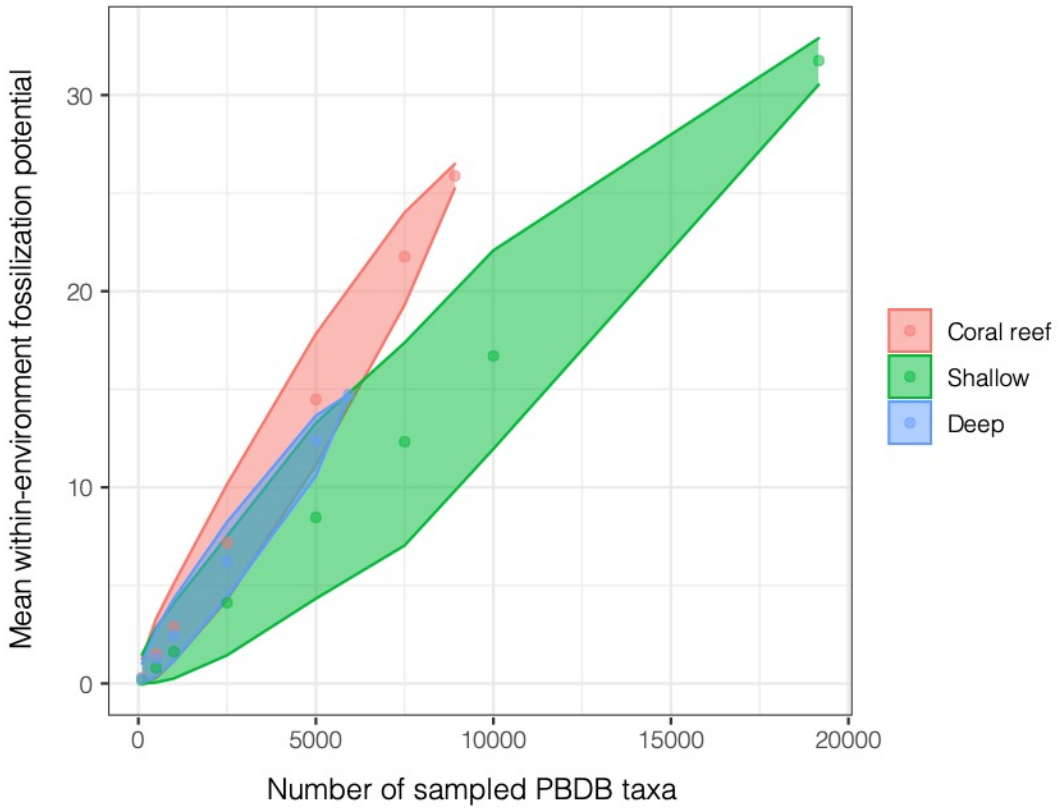
313 **Supplementary Figure S12:** Assemblage-specific genus richness for assemblages with
314 environment (n = 9,669) and substrate (n = 7,545) data plotted against fossilization potential.



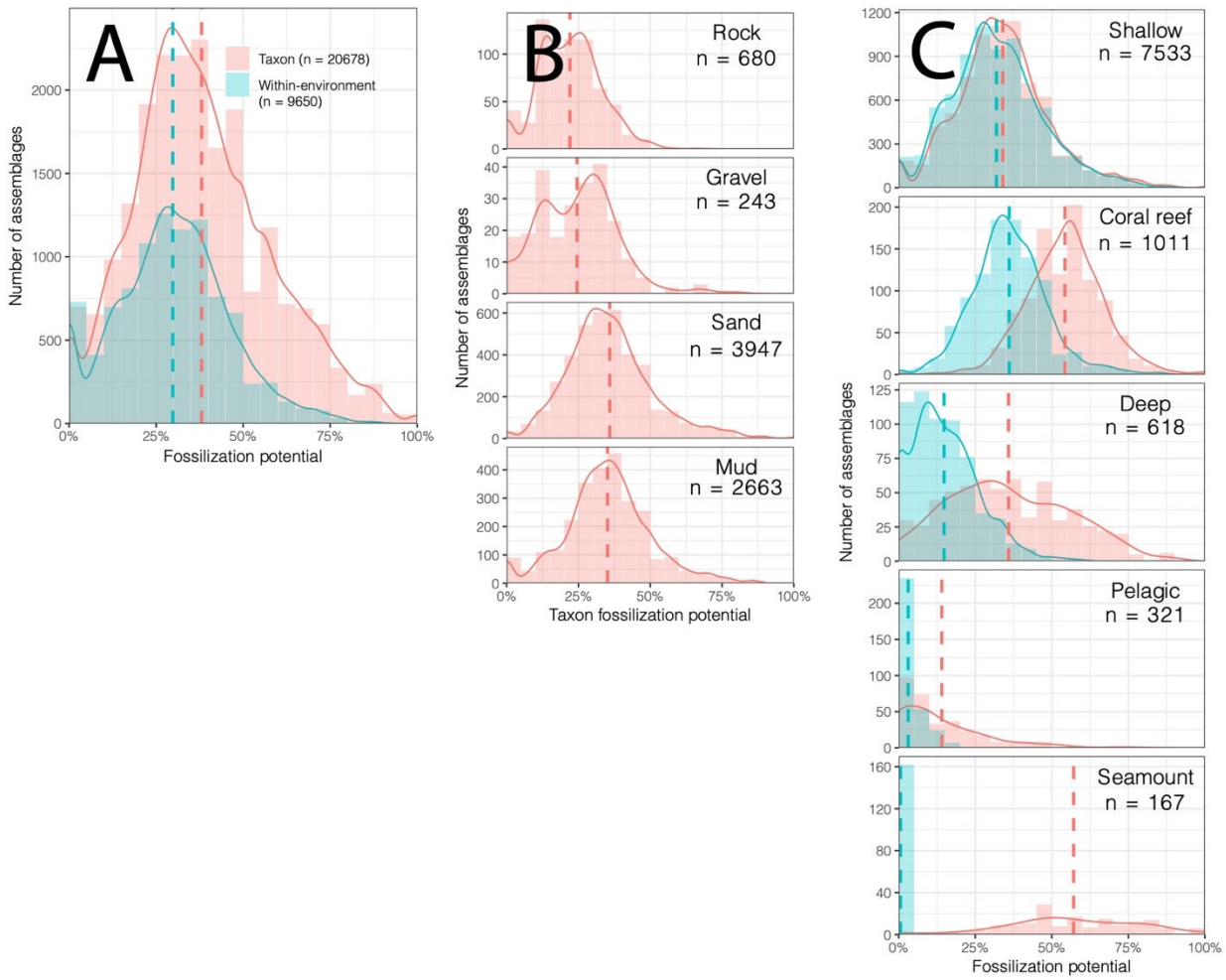
315

316 **Supplementary Figure S13:** Taxon and within-environment fossilization potential mean values
317 for subsampled data (repeated 1000 times). Each mean represents comparisons utilizing random
318 samples of PBDB taxa (5927 per environment, as per the smallest group, deep water) and
319 random samples of OBIS assemblages (1012 per environment, as per the smallest group, coral
320 reefs). Means of the raw data (e.g., Fig. 3) shown using black asterisks. Within-environment
321 fossilization potential for shallow and coral reef environments are off the axes (indicated by
322 circles) and have means of 32% and 26%, respectively.

323



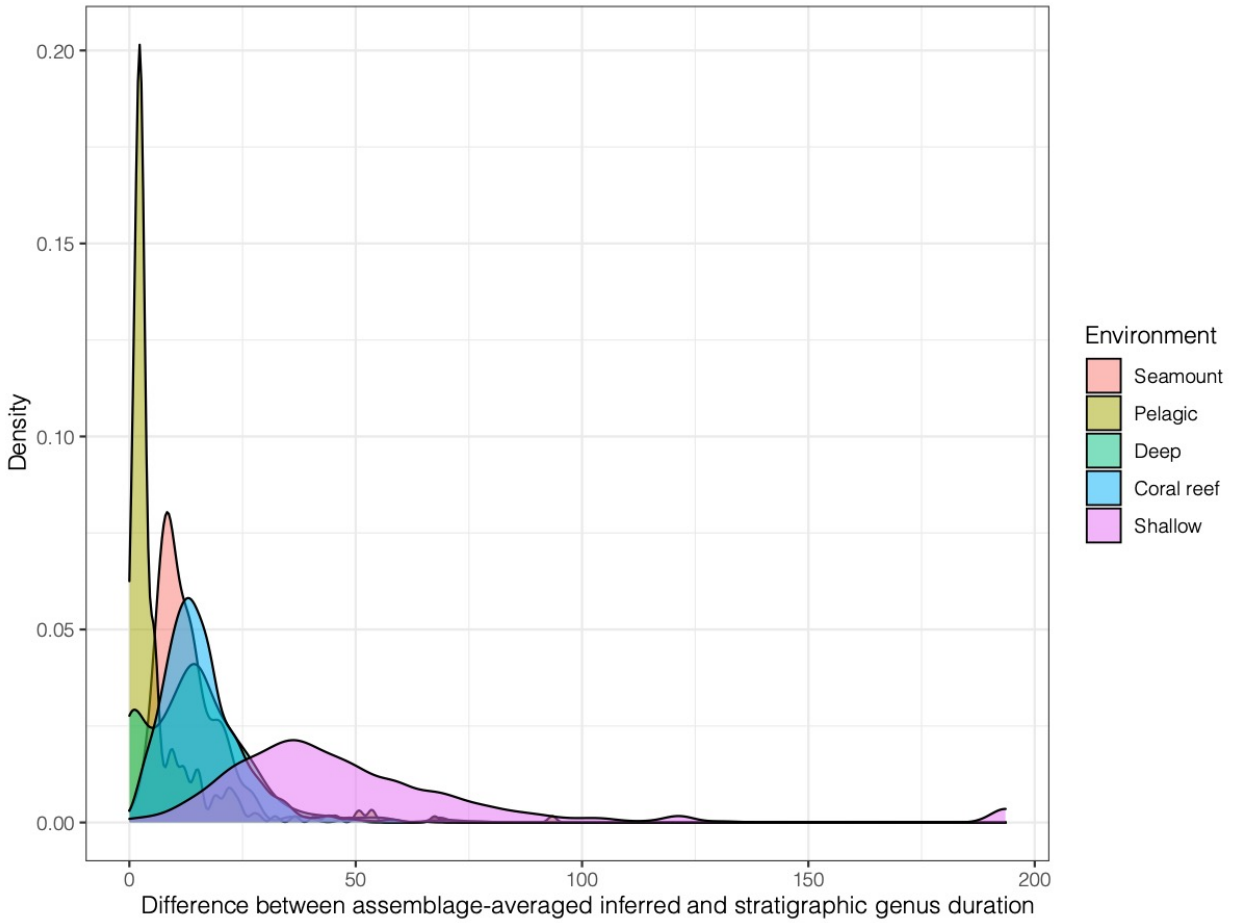
Supplementary Figure S14: 95% confidence intervals of mean within-environment fossilization potential values when calculated based on subsampled PBDB data. Mean within-environment fossilization potential values were generated by first subsampling the number of PBDB taxa per environment (x-axis), then comparing the number of OBIS taxa in an assemblage to the subsampled PBDB faunal list to generate the assemblage-specific fossilization potential value, and finally calculating the mean of all assemblage-specific fossilization potentials by environment. This process was repeated 1000 times at a number of pre-defined PBDB sample sizes (100, 500, 1000, 2500, 5000, 7500, 10000, 20000) up to the maximum number of PBDB taxa assigned to each environment.



335

336 **Supplementary Figure S15:** Invertebrate only analyses. (A) Taxon fossilization potential and
337 within-environment fossilization potential (mean taxon/within-environment fossilization
338 potential = 39/30%). (B) Taxon fossilization potential distributions for shallow water
339 assemblages with substrate information (mean fossilization potential values: rock = 22%, gravel
340 = 25%, sand = 36%, mud = 35%). (C) Taxon and within-environment fossilization potential
341 distributions for assemblages with environmental information (mean taxon/within-environment
342 fossilization potential values: shallow = 33/30%, coral reef = 53/35%, deep = 35/14%, pelagic =
343 14/2%, seamount = 57/0%). Histogram bin widths = 5%; dashed lines indicate mean.

344



345

346 **Supplementary Figure S16:** Density plot of differences between assemblage averaged inferred-
347 and PBDB-durations parsed by environment.

SUPPLEMENTARY DATA DESCRIPTIONS

Supplementary data are available for download at the Harvard Dataverse Repository (https://doi.org/10.7910/DVN/MMMTYZ). Supplementary Data labels and descriptions apply to data files listed in the Dataverse Repository.

Supplementary Data DR1: Raw list of all occurrences in OBIS downloaded on 21st January 2020. Occurrences were culled (procedures in text) using R code (Sup. Data DR4A). Column headers following Darwin Core format (Wieczorek et al., 2012).

Supplementary Data DR2: Raw list of OBIS datasets (downloaded 20th January 2020, Sup. Data 2A) and list of OBIS assemblages (Sup. Data DR2B) meeting the culling procedures described herein and applied using R Code (Sup. Data DR4A).

To generate assemblages, we removed occurrences from Sup. Data 1 (i) lacking data on genus, latitude, longitude, depth, or duration of sampling, (ii) from non-marine settings, or (iii) with inadequate error control (e.g., citizen science data). Occurrences were grouped into assemblages if they were obtained within 0.1° latitude and 0.1° longitude of each other, in the same 10 m depth bin, during the same year, and from the same dataset. We retained only assemblages with representatives of at least three phyla, four classes, five orders, six families, and seven genera in order to avoid poorly sampled assemblages. Alternative binning procedures were tested and did not affect interpretations (Figs. S1).

Additional assemblage data (id_habitat, id_scope, id_plankton, id_sampling, id_fishery, and id_environment) in Sup. Data DR2B was manually assigned to datasets using information from dataset titles and abstracts in Sup. Data DR2A. Column headers in Sup. Data DR2B are as follows: “dataset.par” = unique depth-, year-, location-, and dataset-specific assemblage identifier; “dataset_id” = an identifier for the set of data, which may be a global unique identifier or an identifier specific to a collection or institution; “depth.slice” = 10m depth bin of assemblage; “LongBin” = 0.1° longitudinal bin of assemblage; “LatBin” = 0.1° latitudinal bin of assemblage; “id_habitat” = marine realm of dataset, pelagic or benthic; “id_scope” = taxonomic scope of dataset; “id_plankton” = plankton-only dataset; “id_sampling” = sampling type of assemblage; “id_environment” = environment type of dataset; “title” = name of dataset; “abstract” = abstract of dataset.

Supplementary Data DR3: List of all occurrences in PBDB with select columns, downloaded on 17th January 2020 (A). List of Lagerstätten names generated by manually searching for names in PBDB ("formation") based on the list of Lagerstätten in Muscente et al. (2017) (B).

Data downloaded using the following URL:
http://paleobiodb.org/data1.2/occs/list.csv?datainfo&rowcount&base_name=Animalia&show=full,genus,env

Supplementary Data DR4: R Code used for data cleaning (A), identifying assemblages and calculating fossilization potential (B), analyzing fossilization potential (C), generating genus duration estimates (D; Wang et al., 2016), applying genus duration estimate calculations to OBIS taxa (E), and predicting fossilization potential using conditional inference tree models (F).

Supplementary Data DR5: Seafloor substrate data (Jenkins, 2008; A), code used to create predictive model (B), and predictive model (native R format, ".RDS") applied to OBIS assemblages to assess substrate type (C).

Substrate data (Sup. Data DR5A) used for interpolation were supplied by Jenkins (personal communications): "wd_vals" = water depth; "lat" = latitude, to two decimal places; "lon" = longitude, to two decimal places; "substrate" = dominant substrate type (rock, gravel, sand, and mud).

We interpolated substrate data based on the dbSeabed dataset (Jenkins, 2008) using a k-nearest neighbors (KNN) algorithm. The KNN algorithm is a non-parametric supervised machine learning algorithm used for classification and regression. The KNN predictive model was built using the R package KKNN (Schliep, Hechenbichler, and Lizee, 2016). We performed leave-one-out cross-validation for computationally efficient training of the predictive model. We tested K-values ranging 1-20 and all possible kernel types ("rectangular", "triangular", "epanechnikov", "gaussian", "rank", and "optimal"). For training, dbSeabed data was split so that 80% of the data was used to train the model and 20% to test it. Training identified a K-value of 12 and a triangular kernel as the best parameters, although kernel type did not alter resultant classifications. Our best model had a k-value of 12 and a triangular kernel shape, meaning that

404 the 12-nearest neighbors within a triangle shaped space of the focal grid cell were used to infer
405 the substrate type assigned to that grid cell.

406 Prior to analysis, we extended the dbSeabed data by 10 degrees of longitude on either
407 side (i.e., so that the map spanned -190 to 190 degrees). This allowed the predictive model to
408 utilize data at the edge of its conventional range (-180 to 180 degrees). If this correction was not
409 applied, the model would not know that data at -179 degrees longitude and 179 degrees longitude
410 are proximal. To extend the map, we simply duplicated data in the range of -180 to -170 degrees
411 and 170 to 180 degrees by adding and subtracting, respectively, 360 degrees (i.e., -179 became
412 181 degrees longitude).

413 **Supplementary Data DR6:** PBDB environments matched to broader categories for comparison
414 with OBIS designations.

415 **Supplementary Data DR7:** Tests of sampling biases for shallow water, coral reef, and deep
416 water assemblages (see supplementary information section “Sampling biases” for more
417 information).

REFERENCES

Briggs, D.E.G., Erwin, D.H., and Collier, F.J., 1994, The fossils of Burgess Shale: Washington DC, Smithsonian Institution Press.

Crampton, J.S., Cooper, R.A., Beu, A.G., Foote, M., and Marshall, B.A., 2010, Biotic influences on species duration: interactions between traits in marine molluscs: Paleobiology, v. 36, p. 204–223, doi:https://doi.org/10.1666/09010.1.

Fortey, R.A., 1980, Generic longevity in lower ordovician trilobites: Relation to environment: Paleobiology, v. 6, p. 24–31, doi:https://doi.org/10.1017/S009483730001246X.

Horton, T. et al., 2020, World Register of Marine Species (WoRMS):, https://www.marinespecies.org (accessed January 2020).

Hothorn, T., and Zeileis, A., 2015, partykit : A Toolkit for Recursive Partytioning: Journal of Machine Learning Research, v. 16, p. 3905–3909.

Hou, X., Siveter, D.J., Siveter, D.J., Aldridge, R.J., Cong, P., Gabbott, S.E., Ma, X., Purnell, M.A., and Williams, M., 2017, The Cambrian Fossils of Chengjiang, China: John Wiley & Sons, doi:https://doi.org/10.1002/9781118896372.

Jablonski, D., 2008, Species Selection: Theory and Data: Annual Review of Ecology, Evolution, and Systematics, v. 39, p. 501–524, doi:https://doi.org/10.1146/annurev.ecolsys.39.110707.173510.

Jablonski, D., and Bottjer, D.J., 1983, Soft-bottom epifaunal suspension-feeding assemblages in the Late Cretaceous: implications for the evolution of benthic paleocommunities., *in* Biotic interactions in Recent and fossil benthic communities, p. 747–812.

Jablonski, D., and Hunt, G., 2006, Larval ecology, geographic range, and species survivorship in cretaceous mollusks: Organismic versus species-level explanations: American Naturalist, v. 168, p. 556–564, doi:https://doi.org/10.1086/507994.

Jenkins, C., 2008, dbSEABED: an information processing system for marine substrates:, https://instaar.colorado.edu/~jenkinsc/dbseabed/ (accessed April 2020).

Kammer, T.W., Baumiller, T.K., and Ausich, W.I., 1997, Species longevity as a function of niche breadth: Evidence from fossil crinoids: Geology, v. 25, p. 219–222, doi:https://doi.org/10.1130/0091-7613(1997)025<0219:SLAAFO>2.3.CO;2.

Muscente, A.D. et al., 2017, Exceptionally preserved fossil assemblages through geologic time and space: Gondwana Research, v. 48, p. 164–188, doi:https://doi.org/10.1016/J.GR.2017.04.020.

Schliep, K., Hechenbichler, K., and Lizee, A., 2016, kknn: Weighted k-Nearest Neighbors:, https://cran.r-project.org/web/packages/kknn/index.html (accessed April 2020).

Wang, S.C., Everson, P.J., Zhou, H.J., Park, D., and Chudzicki, D.J., 2016, Adaptive credible intervals on stratigraphic ranges when recovery potential is unknown: Paleobiology, v. 42, p. 240–256, doi:https://doi.org/10.1017/pab.2015.37.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., and Vieglais, D., 2012, Darwin core: An evolving community-developed biodiversity data standard: PLoS ONE, doi:https://doi.org/10.1371/journal.pone.0029715.