Supplementary Materials for

Low-Latitude Origins of the Four Phanerozoic Evolutionary Faunas

A. Rojas*, J. Calatayud, M. Kowalewski, M. Neuman, and M. Rosvall

*Correspondence to: alexis.rojas-briceno@umu.se

This PDF file includes:

Other Supplementary Materials for this manuscript include the following:

Materials and Methods

Data

Genus-level occurrences derive from the Paleobiology Database (PaleoDB; https://paleobiodb.org) (*18*), which at the time of access consisted of 79,976 fossil collections with 448,335 occurrences from 18,297 genera. Here we only included resolved fossil occurrences. The downloaded taxa comprise the well-preserved benthic marine invertebrates (*17*): Brachiopoda, Bivalvia, Gastropoda, Bryozoa, Echinodermata, Anthozoa, Decapoda, and Trilobita. The Paleobiology database assigned fossil collections to paleogeographic coordinates based on their present-day geographic coordinates and geologic age using rotation models provided by the GPlates (http://www.gplates.org). We plotted the geographic maps of the spatial grid cells with the corresponding plate tectonic configuration from GPlates (*31*). Using the *Hexbin* R-package (*32*), we aggregated fossil occurrences into a regular grid of hexagons covering the Earth's surface per each stage in the geological timescale (4,906 grid cells with count > 0; inner diameter = 10° latitude-longitude) (Fig. S1A). This hexagonal binning procedure provides symmetry of neighbors that is lacking in rectangular grids and captures the irregular shape of geographic regions more naturally (*33*). The grid size is a compromise between the lack of spatial resolution provided by hexagons with inner diameter > 10° and an increased number of hexagons with none count when shortening the inner diameter. Nevertheless, study cases on modern marine faunas have demonstrated that network-based biogeographic analyses are robust to the shape (square and hexagonal), size (5° to 10° latitude-longitude), and coordinate system (geographic and projected) of the grid used to aggregate data (*34*, *35*).

Network analysis

We used aggregated occurrence data to generate a multilayer bipartite network (*21*), where layers represent ordered geological stages in the geological timescale (*19*), and two types of nodes in each layer represent taxa and spatiotemporal grid cells (*20*) (Fig. S1). Whereas each taxon can be present in multiple layers, each grid cell is only present in a single layer. To capture interdependencies in the occurrence data in a statistically sound way, we linked taxa to spatiotemporal grid cells through links with weights (w) adjusted for sampling effort. Specifically, for the adjusted weight (w_{ki}) between grid cell k and taxa i, we divided the number of collections at grid cell k that register taxa i by the total number of collections recorded at grid cell k. A similar sampling correction has been employed on previous network-based biogeographic analysis using weighted projections from bipartite occurrence networks (*17, 20*). In addition, we combined the last two Cambrian stages, i.e., Jiangshanian Stage (494 to 489.5 Ma) and Stage 10 (489.5 to 485.4 Ma), into a single layer to account for the lack of data from the younger Stage 10 and to maintain an ordered sequence in the multilayer network framework (*21*). Even though such a gap was placed at the end of the Cambian Period, most grid cells and species from the combined Jiangshanian/Stage 10 (494-485.4 Ma) layer clustered into the Paleozoic supermodule (see below). The assembled multilayer network of the Phanerozoic benthic marine faunas comprises 23,203 nodes (n), including 4,906 spatiotemporal grid cells and 18,297 genera, joined by 144,754 links (m), distributed into 99 layers (t) (Data S1).

To identify important dynamical patterns in the spatiotemporal organization of the Phanerozoic benthic marine faunas as represented in the assembled multilayer network, we used a network clustering approach (Fig. S1B). The conventional approach to partition bipartite occurrence networks based on aggregated fossil occurrences applies standard community

detection methods to the one-mode projection of the original network (*20*). Although such a procedure can provide some insights about the biogeographic structure of ancient marine faunas (*17*), it destroys relevant information regarding higher-order interdependences between taxa and geographic regions. Instead, here we used the map equation multilayer framework (www.mapequation.org), which can operate directly on the multilayer bipartite network and thereby preserve higher-order interdependencies. The map equation multilayer framework consists of an objective function that measures the quality of a given network partition, the map equation itself (*23*), and Infomap, an efficient search algorithm that optimizes this function over different solutions (*21*). We used this method because it can handle bipartite, weighted, and multilayer networks and because it is known for its high performance (*36*-*38*). In addition, Infomap directly provides the number of hierarchical levels within each layer and thus removes the subjectivity inherent in other approaches (*38*).

To capture interdependencies beyond nearest neighbors in the assembled network, the map equation models a random walk on the nodes within and also across layers (Fig. S1B): With probability $(1 - r)$, a random walker moves between taxa and grid cells guided by the weighted intralayer links within its current geological stage, and with probability r, it moves between taxa and grid cells guided by the weighted links in its current geological stage and also in the adjacent geological stages. By relaxing the constraint to allow movement within layers in this way, the multilayer framework enables coupling between adjacent layers such that it accounts for the temporal ordering of geological stages. Consequently, the random walker tends to spend extended times in multilayer modules of strongly connected taxa and grid cells across geological stages. Infomap can identify these modules because using modules in which the random walker persists for relatively long periods optimizes the map equation, which measures how much a modular partition of the nodes can compress a description of the random walker on the network. Following previous network studies, we used the relax rate $r = 0.25$, which is large enough to enable interlayer interdependencies but small enough to preserve intralayer information (*38*). We tested the robustness to the selected relax rate by clustering the assembled network for a range of relax rates and comparing each solution to the solution for $r = 0.25$ using the Jaccard Similarity. Finally, we obtained the reference solution (Data S2) using the assembled network and the following Infomap arguments: `-N 200 -i multilayer --multilayer-relax-rate 0.25 --multilayer-relax-limit 1`. The relax limit is the number of adjacent layers in each direction to which a random walker can move. Thus, a value of 1 enables the temporal ordering of geological stages in the multilayer framework.

We employed a parametric bootstrap for estimating the significance of the multilayer modules delineated in the reference solution. This approach assumes that the assembled network accurately captures connections between benthic taxa and grid cells but that there can be uncertainty in the strength of those interdependencies from variations in sampling effort through time and across space. We resampled taxon occurrence using a truncated Poisson distribution with mean equal to the number of taxon occurrences. The truncated distribution has all probability mass between one and the total number of collections in the grid cell, thus avoiding false negatives. We obtained the resampled link weight by dividing the sampled number by the total number of recorded collections. Using Infomap with the arguments detailed above, we clustered these bootstrapped networks and then compared the resulting partitions with the reference solution. Specifically, for each reference module, we computed the proportion of

bootstrapped partitions where we could find a module with Jaccard similarity higher than 0.5 (P_{05}) and 0.7 (P_{07}) (Tables S1-S2). In addition, we computed the average probability (median) of belonging to a supermodule for nodes of the same layer (Fig. S6). This procedure for estimating module significance is described in (39), which includes a case study on biogeographic networks of modern vertebrates.

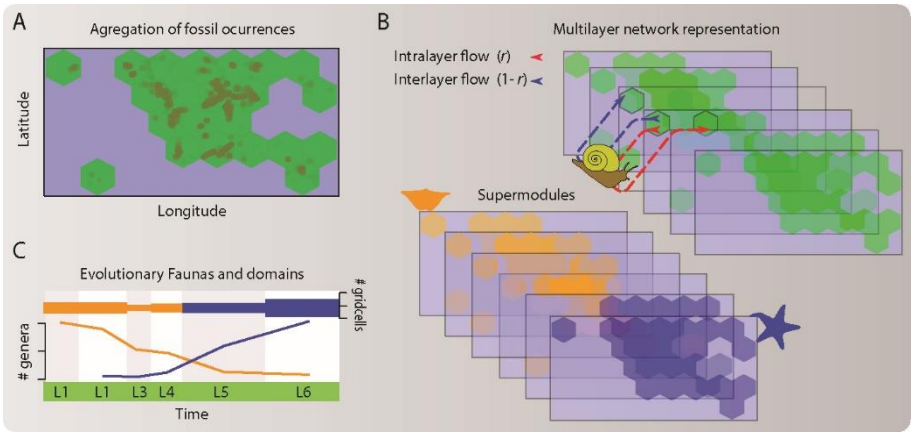

Figure. S1. Multilayer network representation of global fossil occurrences and visualization of its modular structure.

(A) Data aggregation. We aggregated global-scale fossil occurrences into hexagonal spatial grid cells. (B) Network representation and clustering. We constructed a multilayer network representation (*21*) of the aggregated data by joining taxa to grid cells in each stage (L1 to L6) through links adjusted for sampling effort and layers representing ordered geological stages. We used the hierarchical network clustering algorithm called Infomap (*22*) to delineate groups of highly interconnected taxa and grid cells across layers with multilayer modules. (C) Mapping evolutionary faunas and domains. We mapped faunas and temporal domains using the chronostratigraphic distribution of the module grid cells and per-layer taxa richness.

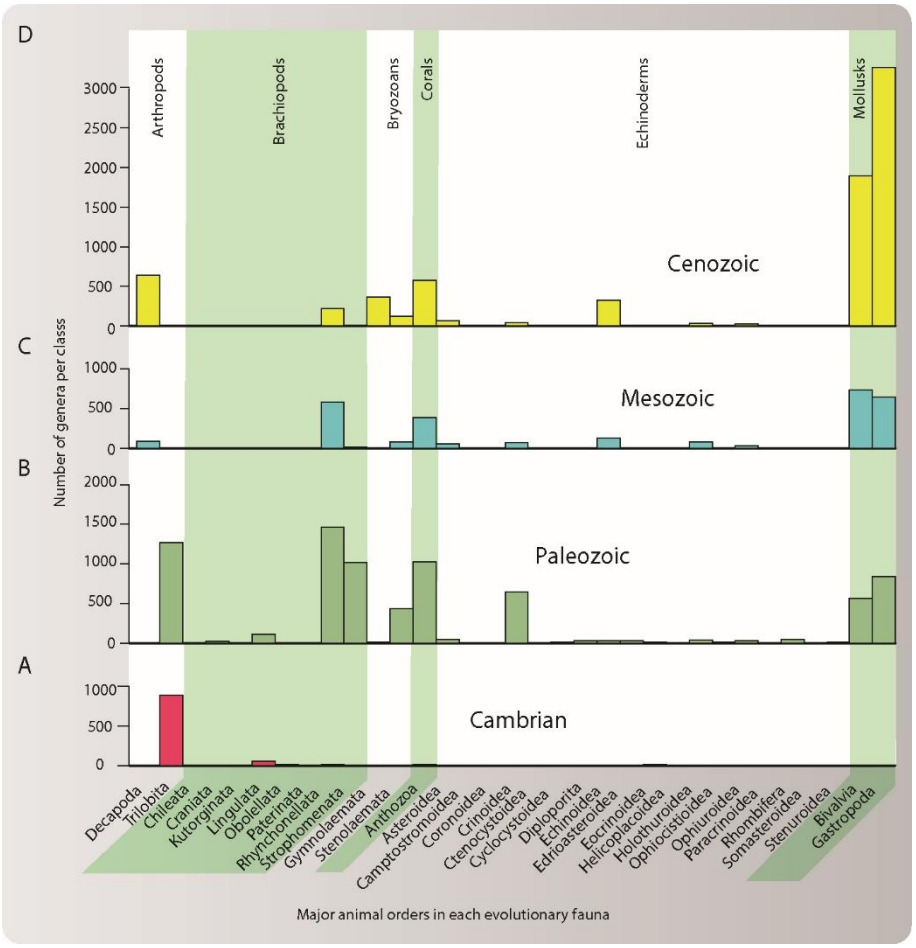

Figure. S2. Class-level composition of the four marine evolutionary faunas.

Clustered taxa define four partially overlapping sets of benthic marine animals. (A) Cambrian evolutionary fauna. (B) Paleozoic evolutionary fauna. (C) Mesozoic evolutionary fauna. (D) Cenozoic evolutionary fauna. The classes of marine invertebrates that contribute the most to the Cambrian, Paleozoic, and combined Paleozoic-Mesozoic mega-assemblages delimited here match those from the *Three Great Evolutionary Faunas [if you decide not to capitalize and italicize this phrase in the manuscript, make that change here also](1*). The Cambrian mega-assemblage comprises trilobites (88%) and lingulates (5%); the Paleozoic domain comprises rhynchonellids (19%), trilobites (16%), anthozoans (13%), strophomenids (13%), gastropods (11%), crinoids (8%), bivalves (7%), and stenolaemate bryozoans (6%); the Mesozoic domain comprises bivalves (25%), gastropods (22%), rhynchonellids (20%), and anthozoans (13%); and the Cenozoic domain comprises gastropods (43%), bivalves (25%), decapods (8%), and anthozoans (8%).

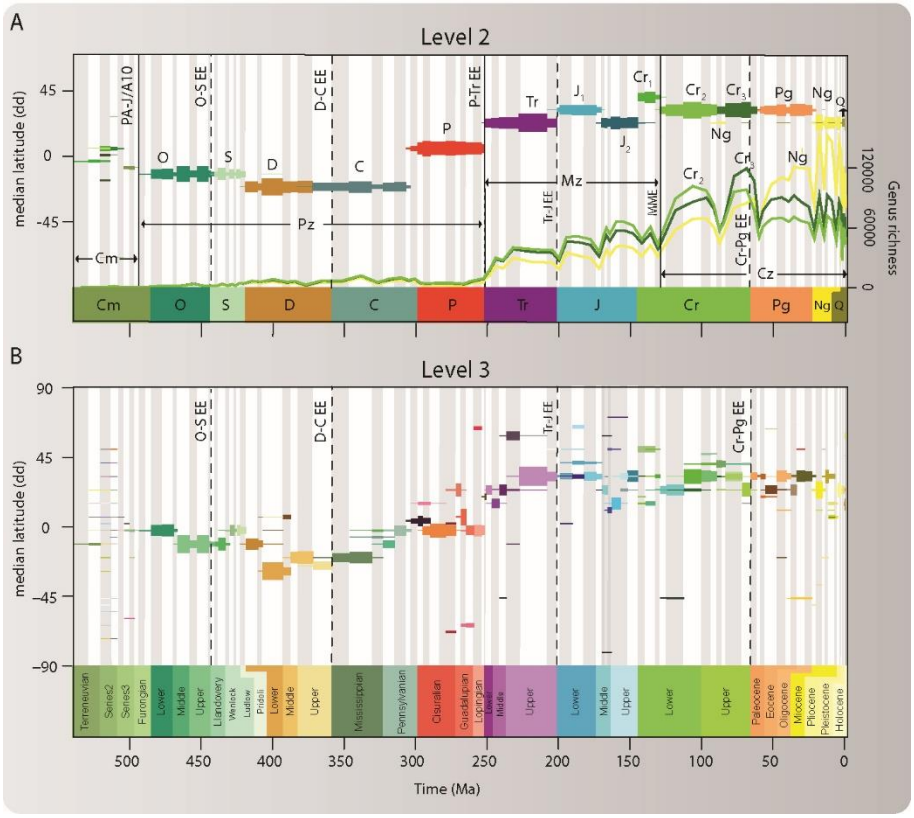

Figure. S3. Lower-level modules in the configuration of the multilayer network of the Phanerozoic benthic marine faunas.

Lower-level modules capture internal structure of the four evolutionary faunas. (A) Second hierarchical level (Level 2). Lines represent the genus richness of the faunas associated with Cretaceous and Neogene modules. Horizontal bars represent the number of module grid cells in each time interval. The Cenozoic fauna consists of Cretaceous (Cr_2 and Cr_3), Paleogene (Pg), Neogene (Ng), and Quaternary (Q) modules (all $P_{0.7} \geq 0.99$). The Mesozoic fauna consists of Triassic (Tr), Jurassic (J_1, J_2) and Cretaceous (Cr_1) modules (all $P_{0.7} \geq 0.98$). The Paleozoic fauna consists of Ordovician, Silurian, Devonian, Carboniferous, and Permian modules (all $P_{0.7} \geq 0.94$). The Cambrian consists of various small modules (*five* modules all $P_{0.7} \geq 0.58$; 4 modules all $P_{0.7} \leq 0.41$). (B) Third hierarchical level (Level-3) (Table S2). Some of these lower-level modules form geographically coherent units underlying the evolutionary faunas.

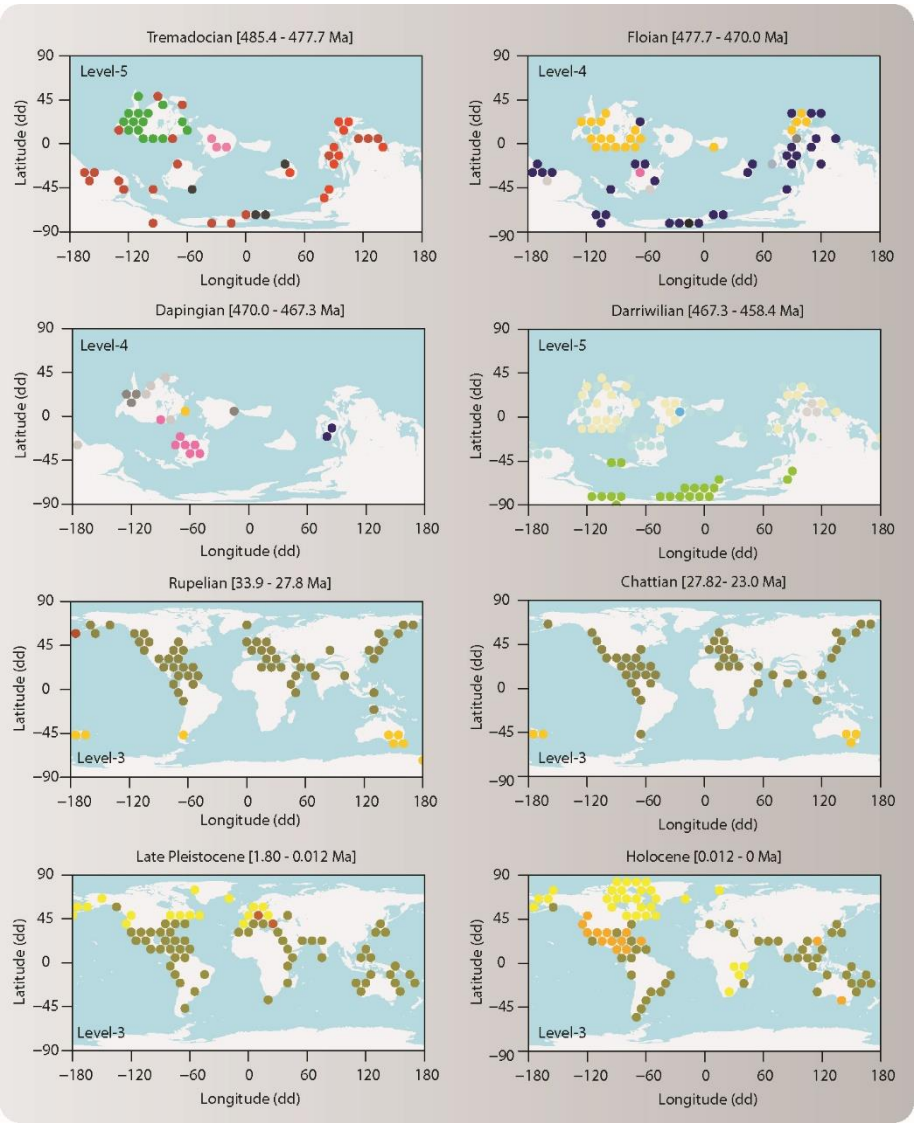

Figure. S4. Examples of marine bioregions underlying the four evolutionary faunas.

Geographic maps of lower-level modules. Circles represent grid cells colored by their module affiliation (Data S2). Lower-level modules form geographically coherent bioregions (*17*, *20*) underlying the evolutionary faunas in the modular organization of the Phanerozoic marine life.

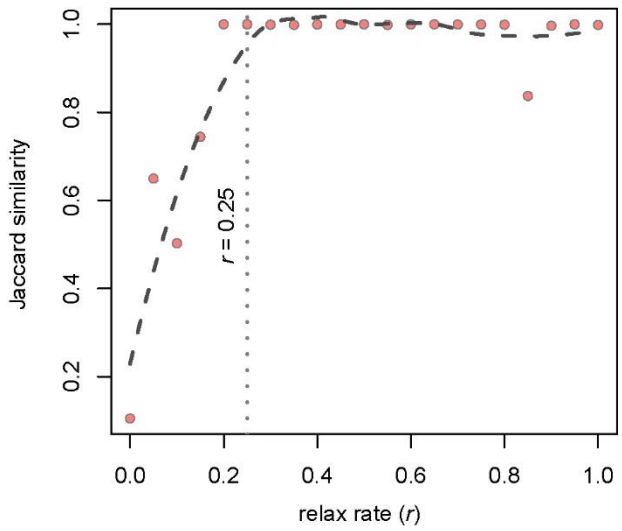

Figure. S5. Network clustering robustness to the selected the relax rate (*r*).

Network clustering results are highly robust to variations in the relax rate (*r*). The plot illustrates the similarity of the reference solution ($r = 0.25$) with*[similarity to?]* solutions obtained from different relax rates. Results are particularly robust in the domain $r \geq 0.20$. Following previous studies on complex networks, we used a relax rate $r = 0.25$ for the reference solution, which is large enough to enable interlayer interdependencies but small enough to preserve intralayer information (*38*).

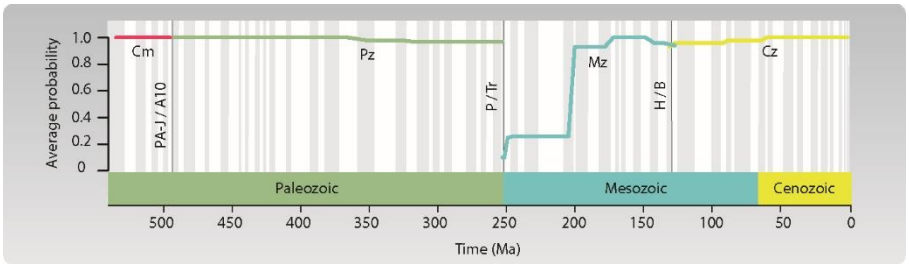

Figure S6. Stage-level significance of the supermodules delineated in the assembled network.

The average probability (median) of belonging to a supermodule for nodes of the same layer was calculated according to (*39*). It shows the instability of the modular structure in the assembled network after the Earth's largest mass extinction event (*6, 24*). This stage-level pattern explains the overall significance ($P_{0.7}$ = 0.25) of the Mesozoic Evolutionary Fauna (Fig. 1, Table S1). Abbreviations: Cambrian (Cm); Paleozoic (Pz); Mesozoic (Mz); and Cenozoic (Cz). Boundaries: combined Paibian-Jiangshanian—Age10 (PA-J/A10); Permian—Triassic (P/Tr); and Hauterivian—Barremian (H/B).

Supplementary Data S1.
Multilayer network of Phanerozoic benthic marine animals in multilayer network format. This standard file specifies nodes and links in two different sections. The first section includes the node indexes and names. The second section describes the intralayer link structure; each row includes layer index, source node index, target node index, and link weight. Interlayer links derive from the intralayer link structure by relaxing the layer constraints on those links with probability *r*.

Supplementary Data S2.

Reference solution in plain text format. This standard file contains the best hierarchical partition (shortest description length) of the attempts. Each row begins with the multilevel module assignments of a node in a colon-separated format and ordered from coarse (supermodules) to fine level. Modules within each hierarchical level are sorted by the total amount of flow they contain – their steady state population of random walkers (*23*). The decimal number is the amount of flow in each node. The last integer corresponds to the index of the node in the multilayer network file (Data S1).

Supplementary Table S1.

Robustness results of the multilayer network analysis of the fossil record of Phanerozoic benthic marine faunas: First (supermodules) and second hierarchical levels (Level-2).

Supplementary Table S2.

Robustness results of the multilayer network analysis of the fossil record of Phanerozoic benthic marine faunas: Third hierarchical level (Level-3).

Additional References

31. N. Wright, S. Zahirovic, R. D. Müller, M. Seton, Towards community-driven paleogeographic reconstructions: integrating open-access paleogeographic and paleobiology data with plate tectonics. *Biogeosciences*. **10**, 1529–1541 (2013).
32. Dan Carr, N, Lewin-Koh, M, Maechler, D, Sarkar (2018). hexbin: Hexagonal Binning Routines. R package version 1.27.2. https://CRAN.R-project.org/package=hexbin (2016).
33. C. P. D. Birch, S. P. Oom, J. A. Beecham, Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological Modelling*. **206**, 347–359 (2007).
34. M. J. Costello et al., Marine biogeographic realms and species endemicity. *Nat Commun*. **8**, 1057 (2017).
35. D. A. Vilhena *et al.*, Bivalve network reveals latitudinal selectivity gradient at the end-Cretaceous mass extinction. *Sci Rep*. **3**, 1790 (2013).
36. A. Lancichinetti, S. Fortunato, Community detection algorithms: A comparative analysis. *Phys. Rev. E*. **80**, 056117 (2009).
37. R. Aldecoa, I. Marín, Exploring the limits of community detection strategies in complex networks. *Sci Rep*. **3**, 2216 (2013).
38. U. Aslak, M. Rosvall, S. Lehmann, Constrained information flows in temporal networks reveal intermittent communities. *Phys. Rev. E*. **97**, 062312 (2018).
39. J. Calatayud, R. Bernardo-Madrid, M. Neuman, A. Rojas, M. Rosvall, Exploring the solution

landscape enables more reliable network community detection. arXiv:1905.11230 [physics] (2019) (available at http://arxiv.org/abs/1905.11230).