Supplementary material to the article

Diatom studies three decades into the molecular age: a bibliometric analysis reveals molecular underexploration of diatoms compared to other taxa

by

Jan Kollár, Kateřina Kopalová, and Tyler J. Kohler

Contents

Note S1. Search queries for the datasets

General structure of the queries:

taxonomic part AND methodological part AND timespan part

Diatoms *whole*:

(TS=(diatom* NOT diatomite* NOT diatomous* NOT diatomic* NOT diatomaceous earth) OR TS=(bacillariophy*)) AND (PY=(1988-2023))

Diatoms *molecular*:

(TS=(diatom* NOT diatomite* NOT diatomous* NOT diatomic* NOT diatomaceous earth) OR TS=(bacillariophy*)) AND (TS=(molecular OR DNA OR rDNA OR RNA OR rRNA OR eDNA OR aDNA OR "gene" OR genetic OR DNA barcod* OR metabarcod* OR SSU OR 16S OR 18S OR LSU OR 28S OR 5S OR rbc* OR cox1 OR COI OR COXI OR psb* OR matK OR microsatellite* OR MSAT* OR AFLP OR RFLP OR *genom* OR *proteom* OR *transcript* OR *metabolom* OR amplicon OR NGS OR next-generation sequencing OR HTS OR high-throughput sequencing OR Sanger sequencing OR nucleotide OR SNP* OR PCR OR polymerase chain reaction OR *exom* OR exon* OR intron*)) AND (PY=(1988-2023))

Diatoms electron microscopy (*em*):

(TS=(diatom* NOT diatomite* NOT diatomous* NOT diatomic* NOT diatomaceous earth) OR TS=(bacillariophy*)) AND (TS=(SEM OR "electron microscop*" OR TEM OR EM OR ultractructur*)) AND (PY=(1988-2023))

Diatoms core diversity, ecology, and evolution (*mol-eco-evo*):

((TI=(diatom* NOT diatomite* NOT diatomous* NOT diatomic* NOT diatomaceous earth) OR TI=(bacillariophy*)) OR (AK=(diatom* NOT diatomite* NOT diatomous* NOT diatomic* NOT diatomaceous earth) OR TI=(bacillariophy*))) AND ((TI=(molecular OR DNA OR rDNA OR RNA OR rRNA OR eDNA OR aDNA OR "gene" OR genetic OR DNA barcod* OR metabarcod* OR SSU OR 16S OR 18S OR LSU OR 28S OR 5S OR rbc* OR cox1 OR COI OR COXI OR psb* OR matK OR microsatellite* OR MSAT* OR AFLP OR RFLP OR *genom* OR *proteom* OR *transcript* OR *metabolom* OR amplicon OR NGS OR next-generation sequencing OR HTS OR high-throughput sequencing OR Sanger sequencing OR nucleotide OR SNP* OR PCR OR polymerase chain reaction OR *exom* OR exon* OR intron*)) OR (AK=(molecular OR DNA OR rDNA OR RNA OR rRNA OR eDNA OR aDNA OR "gene" OR genetic OR DNA barcod* OR metabarcod* OR SSU OR 16S OR 18S OR LSU OR 28S OR 5S OR rbc* OR cox1 OR COI OR COXI OR psb* OR matK OR microsatellite* OR MSAT* OR AFLP OR RFLP OR *genom* OR *proteom* OR *transcript* OR *metabolom* OR amplicon OR NGS OR next-generation sequencing OR HTS OR high-throughput sequencing OR Sanger sequencing OR nucleotide OR SNP* OR

PCR OR polymerase chain reaction OR *exom* OR exon* OR intron*))) AND (PY=(1988-2023)) AND (WC=("Plant Sciences" OR "Evolutionary Biology" OR "Ecology" OR "Biodiversity Conservation" OR "Marine Freshwater Biology" OR "Oceanography" OR "Environmental Sciences" OR "Limnology" OR "Microbiology" OR "Multidisciplinary Sciences" OR "Palaeontology" OR "Biology") NOT WC=("Biotechnology Applied Microbiology"))

The taxon-defining parts:

diatoms - (diatom* NOT diatomite* NOT diatomous* NOT diatomic* NOT diatomaceous earth) OR (bacillariophy*)

cyanobacteria - (cyanobacteria OR "blue-green alga*" OR "blue green alga*" OR cyanophy*)

rhodophytes - (rhodophy* OR "red alga*" OR "red seaweed")

phaeophytes - (phaeophy* OR "brown seaweed" OR "brown alga*")

chrysophytes *s.s.* - chrysophy*

synurophytes - synurophy*

eustigmatophytes - eustigmatophy*

glaucophytes - glaucophy*

euglenophytes - euglenophy*

cryptophytes - (cryptophy* OR cryptomonad*)

dinophytes - (dinophy* OR dinoflagellate*)

chlorophytes *s.s.* - (chlorophyt* OR chlorophyc* OR "green alga*" OR "green seaweed")

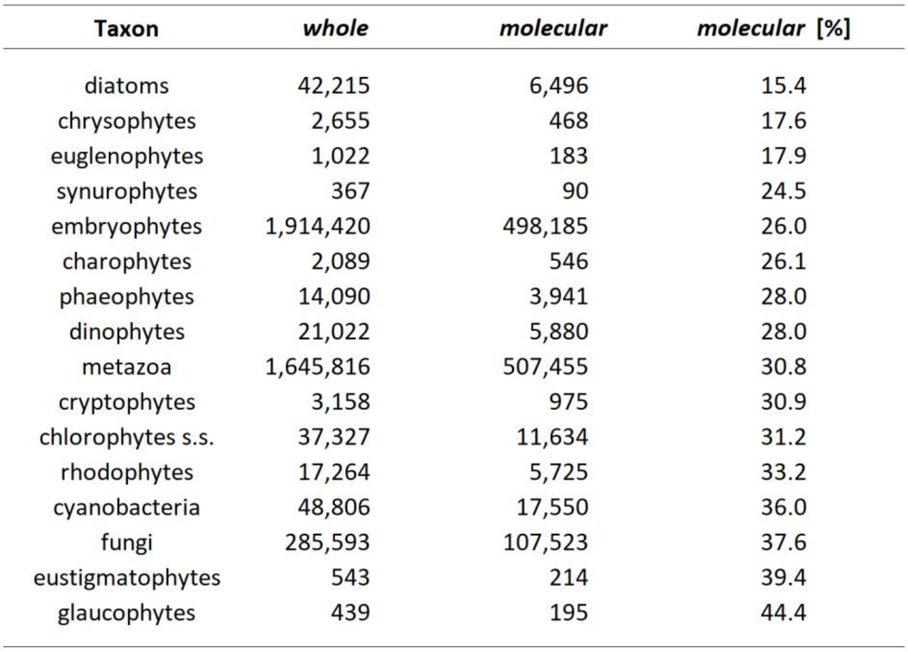
charophytes - charophy*

embryophytes - ((plant* OR angiosperms* OR gymnosperm* OR "vascular plant*" OR bryophy* OR fern* OR magnoliophy* OR archaeplastida OR petridiophy* OR embryophy*) NOT (algae OR seaweed OR glaucophy* OR rhodophy* OR phaeophy* OR chlorophy* OR charophy*))

fungi - (fungi OR mycology OR mycota OR ascomycota OR basidiomycota OR chytridiomycota OR glomeromycota OR Zygomycota)

metazoan - (metazoa OR animals OR fauna OR invertebrates OR vertebrates OR chordata OR arthropoda OR mollusca OR cnidaria OR echinodermata OR platyhelminthes OR nematoda OR annelida OR porifera OR coelenterate)

Table S1. Research published between 1988 and 2023 for different taxa given as number of documents (*whole*), number of documents incorporating molecular methods (*molecular*), and its proportion.



Taxon	*whole*	*molecular*	*molecular* [%]
diatoms	42,215	6,496	15.4
chrysophytes	2,655	468	17.6
euglenophytes	1,022	183	17.9
synurophytes	367	90	24.5
embryophytes	1,914,420	498,185	26.0
charophytes	2,089	546	26.1
phaeophytes	14,090	3,941	28.0
dinophytes	21,022	5,880	28.0
metazoa	1,645,816	507,455	30.8
cryptophytes	3,158	975	30.9
chlorophytes s.s.	37,327	11,634	31.2
rhodophytes	17,264	5,725	33.2
cyanobacteria	48,806	17,550	36.0
fungi	285,593	107,523	37.6
eustigmatophytes	543	214	39.4
glaucophytes	439	195	44.4

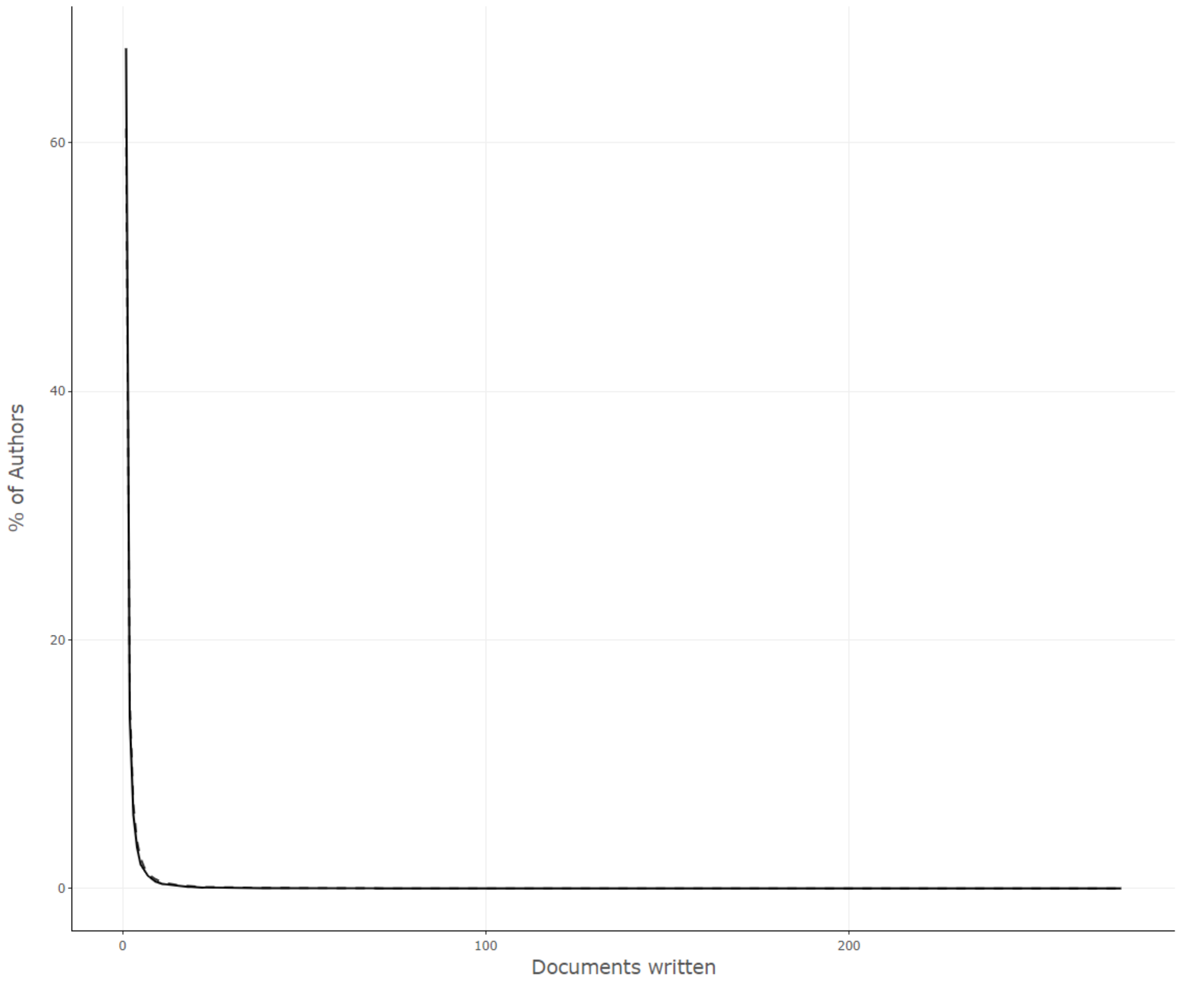


Figure S1. Proportion of authors publishing given number of documents in the whole body of diatom research between 1988 and 2023 (solid line). The data perfectly match the prediction by Lotka's Law (dashed line; Lotka 1926). Only 32.4% of authors published ≥2 documents.

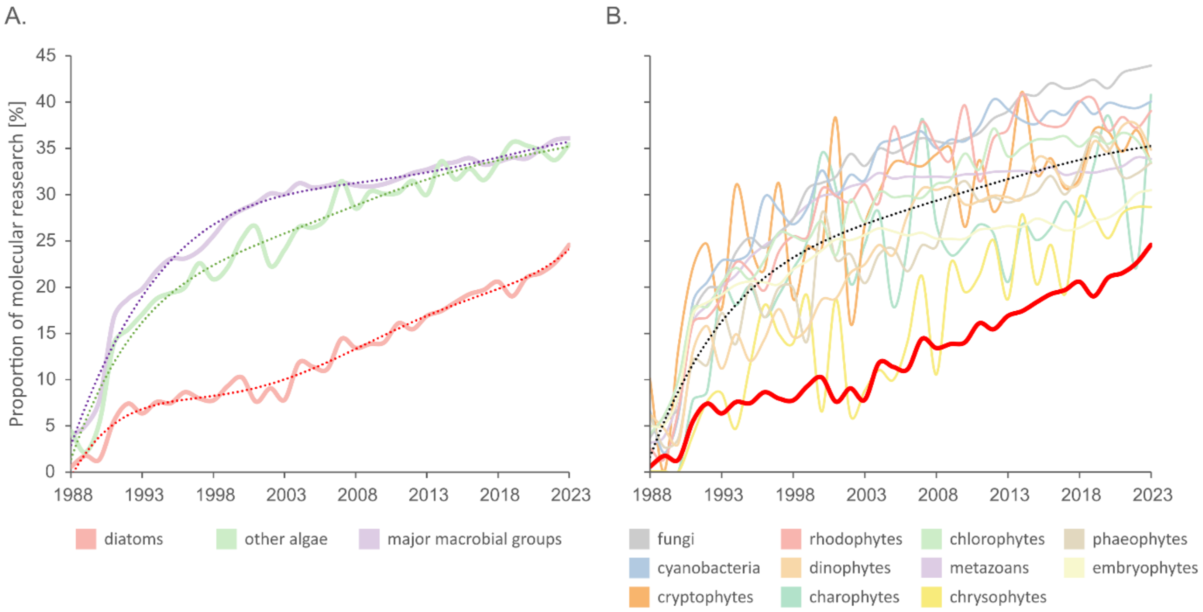


Figure S2. Development of a proportion of molecular research over time between 1988 and 2023 for different taxonomic groups. Panel A compares diatoms (red) to the averages for other algae (including cyanobacteria; green) and major macroscopic groups (i.e., metazoans, embryophytes, and fungi; purple). Solid lines are averages themselves, while the dotted lines represent polynomial functions based on these averages. Panel B is identical to Figure 3, except that taxonomic groups with datasets larger than 2,000 documents are shown (instead of 10,000 documents threshold applied in Figure 3). Diatoms (red line) are compared to other taxa (pastel colours in the background). Chrysophytes are here approached in a broad sense, i.e., merging both chrysophytes *s.s.* and synurophytes. Datasets below 2,000 documents included euglenophytes (1,022 documents), eustigmatophytes (543), and glaucophytes (439). The black dotted line is based on an average for the other-than-diatom taxa shown.

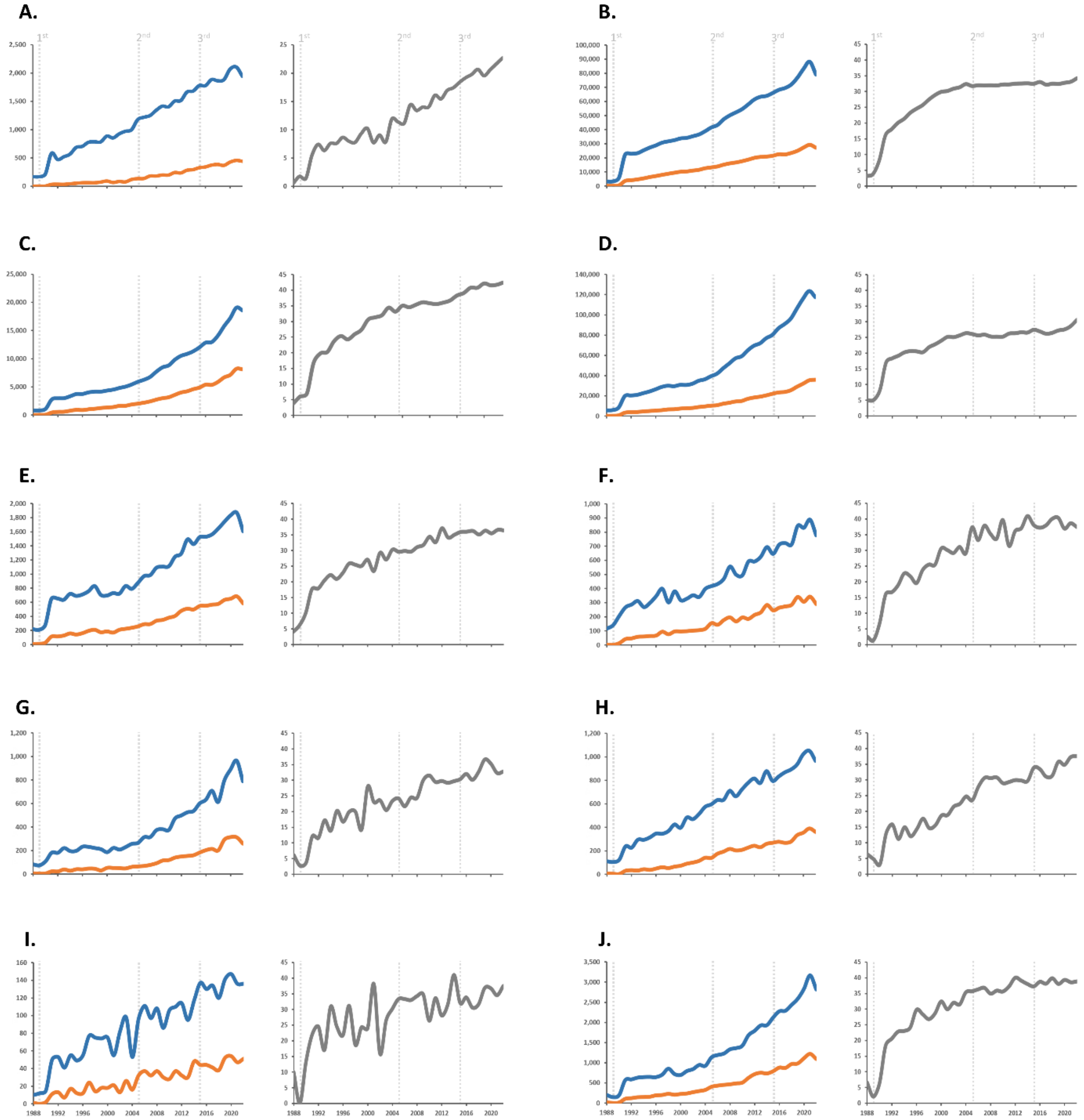


Figure S3. Saturation plots for (A) diatoms, (B) metazoans, (C) fungi, (D) embryophytes, (E) chlorophytes s.s., (F) rhodophytes, (G) phaeophytes, (H) dinophytes, (I) cryptophytes, and (J) cyanobacteria. The plots depict *whole* (blue) and *molecular* (orange) annual production (in numbers of articles) at the top, while an annual proportion of molecular research within given taxonomic group is given at the bottom (grey). Except for diatoms (A), development of the annual proportion of molecular research in all groups seems to be approximately asymptotic, suggesting saturation of the field (e.g., zoology, mycology, botany) by molecular methods. Major milestones in the development of molecular technology are indicated by dotted lines (1st, first commercial PCR thermocycler; 2nd, second generation sequencing platforms commercialized; 3rd, Oxford Nanopore third generation sequencing technology commercialized). Trends of annual production (both *whole* and *molecular*) observed in different parts of the timespan (e.g., exponential increase after 1989, or sudden drop in 2022-2023) are universal across taxa. The drop in 2022 and 2023 is most likely caused by database lag, although some influence of the COVID-19 pandemics cannot be excluded.

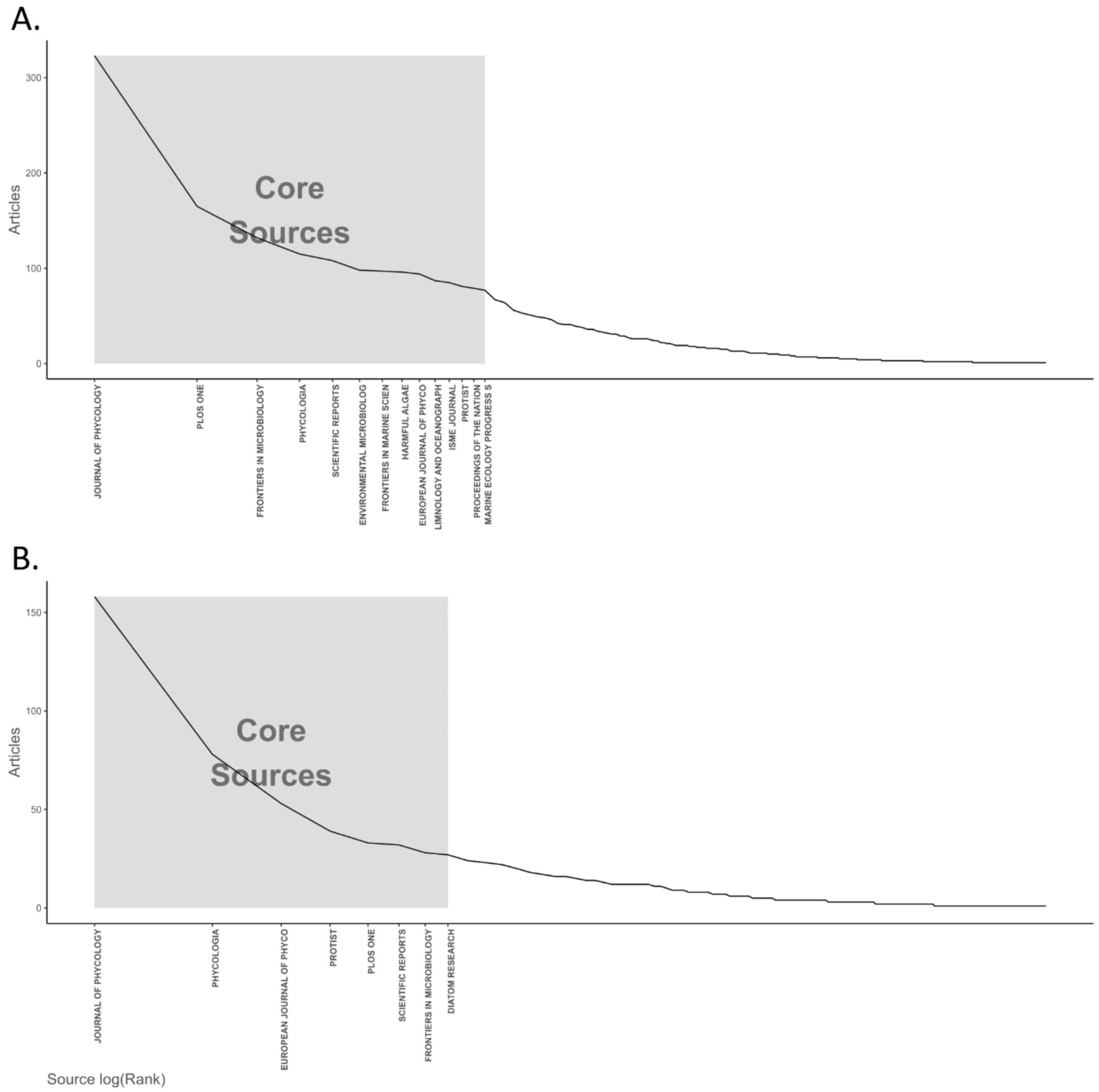


Figure S4. Core journals of molecular diatom research published between 1988 and 2023 (A; the *molecular* dataset, n = 6,496 documents), and its subset broadly focused on diatom diversity, ecology, and evolution (B; the *mol-eco-evo* dataset, n = 1,295), as identified by Bradford's law (see section *Molecular diatomists and where to find them*).

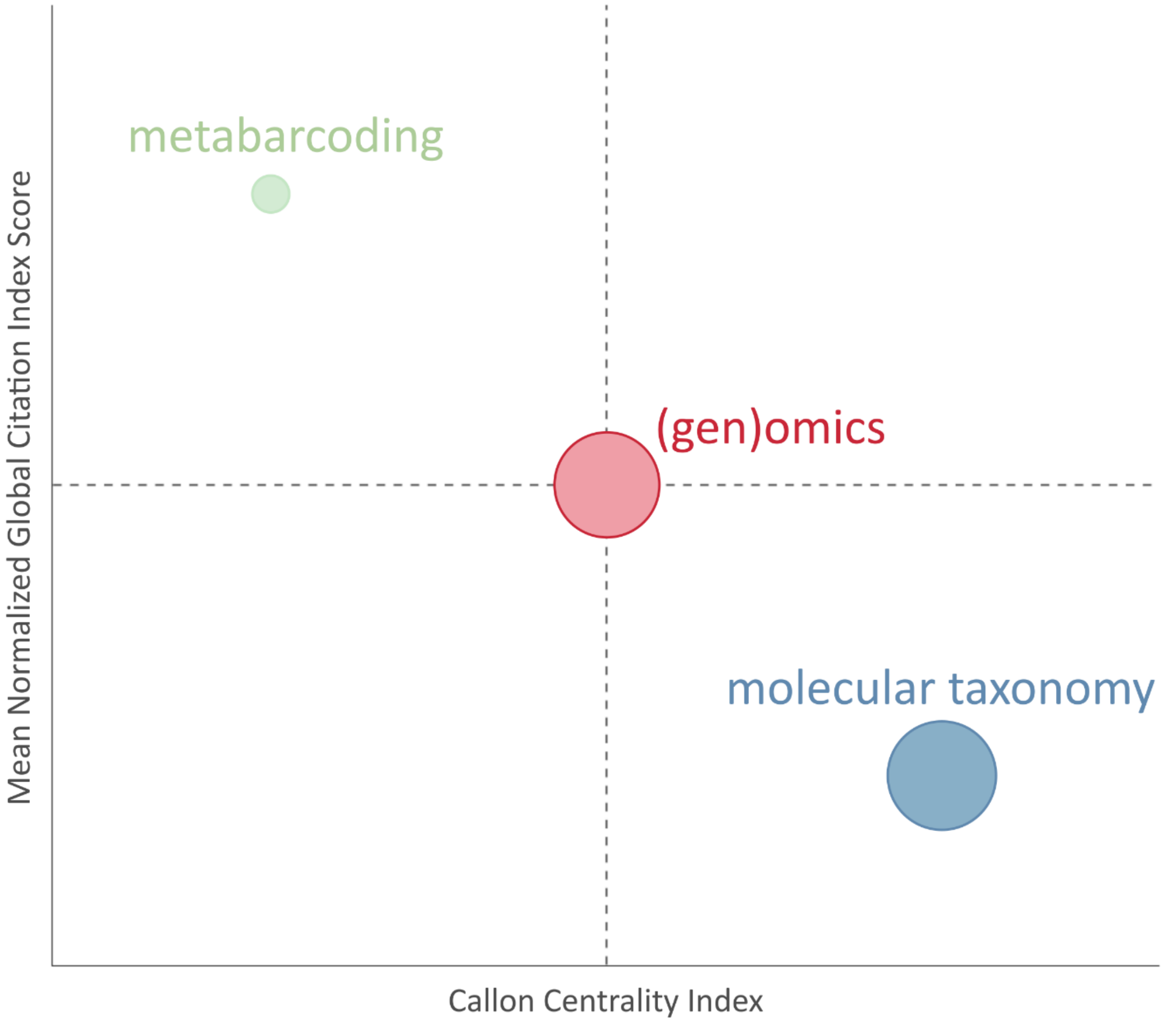


Figure S5. Results of a coupling network analysis depicted as a coupling map. The x-axis represents centrality of the detected clusters (measured as *Callon's Centrality* index; Callon et al. 1991), while the y-axis represents an impact of the clusters (measured as *Mean Normalized Global Citation Score*; Bornmann and Haunschild 2016). The underlaying logic of the analysis is that if both document A and document B cite document C, then A and B are topically related. The analysis is based on the top-cited 1,000 documents of the *mol-eco-evo* dataset (i.e., subset of molecular diatom research broadly focused on diversity, ecology, and evolution). The size of the circles reflects number of documents present within a given cluster. Based on coupling, the analysis identified three separate clusters within the field of molecular diatomology, which we interpreted as (gen)omics (red; contains 44.1% of the documents), molecular taxonomy in a broad sense (blue; 49.2%), and metabarcoding (green; 6.7%). An interested reader may explore our datasets using a relatively user-friendly biblioshiny app (www.bibliometrix.org/home/index.php/layout/biblioshiny).

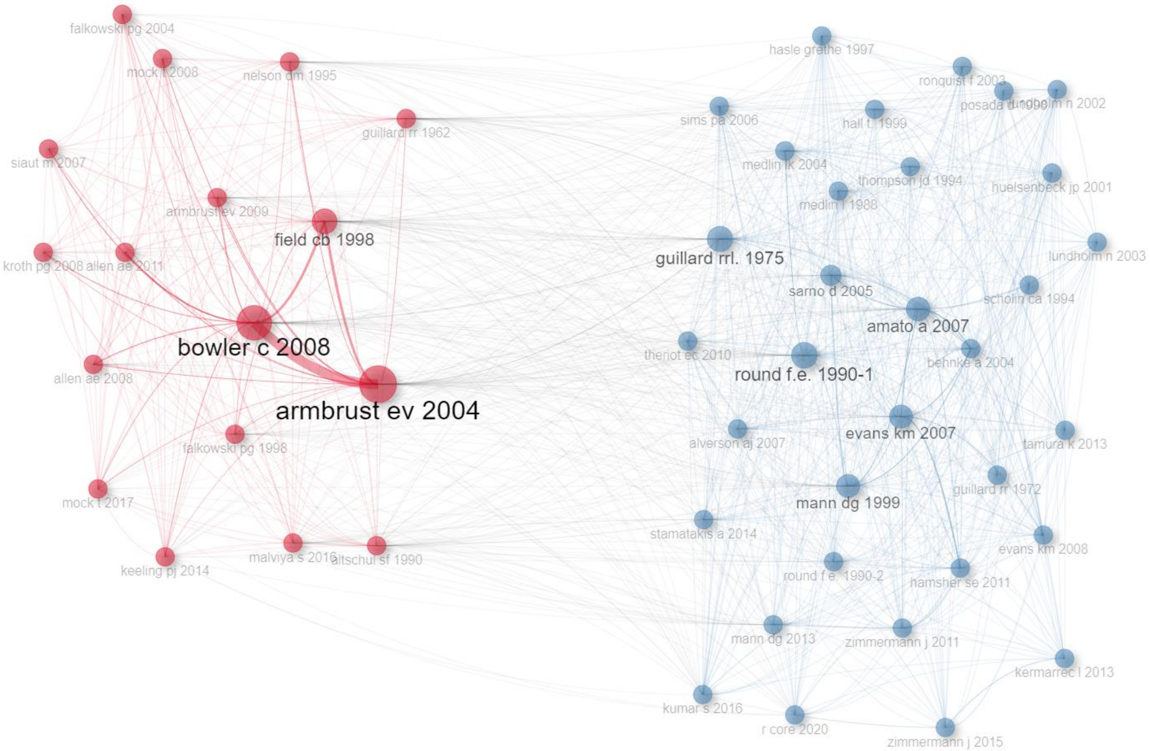


Figure S6. Co-citation network with documents of the *mol-eco-evo* dataset (i.e., subset of molecular diatom research broadly focused on diversity, ecology, and evolution) as the units of analysis. The basic logic of the analysis follows: if both document A and document B are cited by document C, then A and B are topically related. The nodes (i.e., circles) represent documents co-cited by the analysed documents, which is why documents outside of the dataset appear (e.g., 'guillard rr 1962', standing for Guillard and Ryther 1962, *Can. J. Microbiol.* 8). The thickness of the edges (i.e., lines) represent frequency of co-citation of a given pair of documents (e.g., Ambrust et al. 2004, *Science* 306, and Bowler et al. 2008, *Nature* 456, published first and second diatom genomes, and are thus frequently cited together). Only 50 (locally, i.e., within the dataset) top-cited documents were included in the analysis. Sizes of the nodes correspond to the number of local citations. Based on co-citations, the analysis identified two separate clusters within the field of molecular diatomology, which we interpreted as (gen)omics (red), and molecular taxonomy in a broad sense (blue; see section *Conceptual structure of molecular diatomology*).

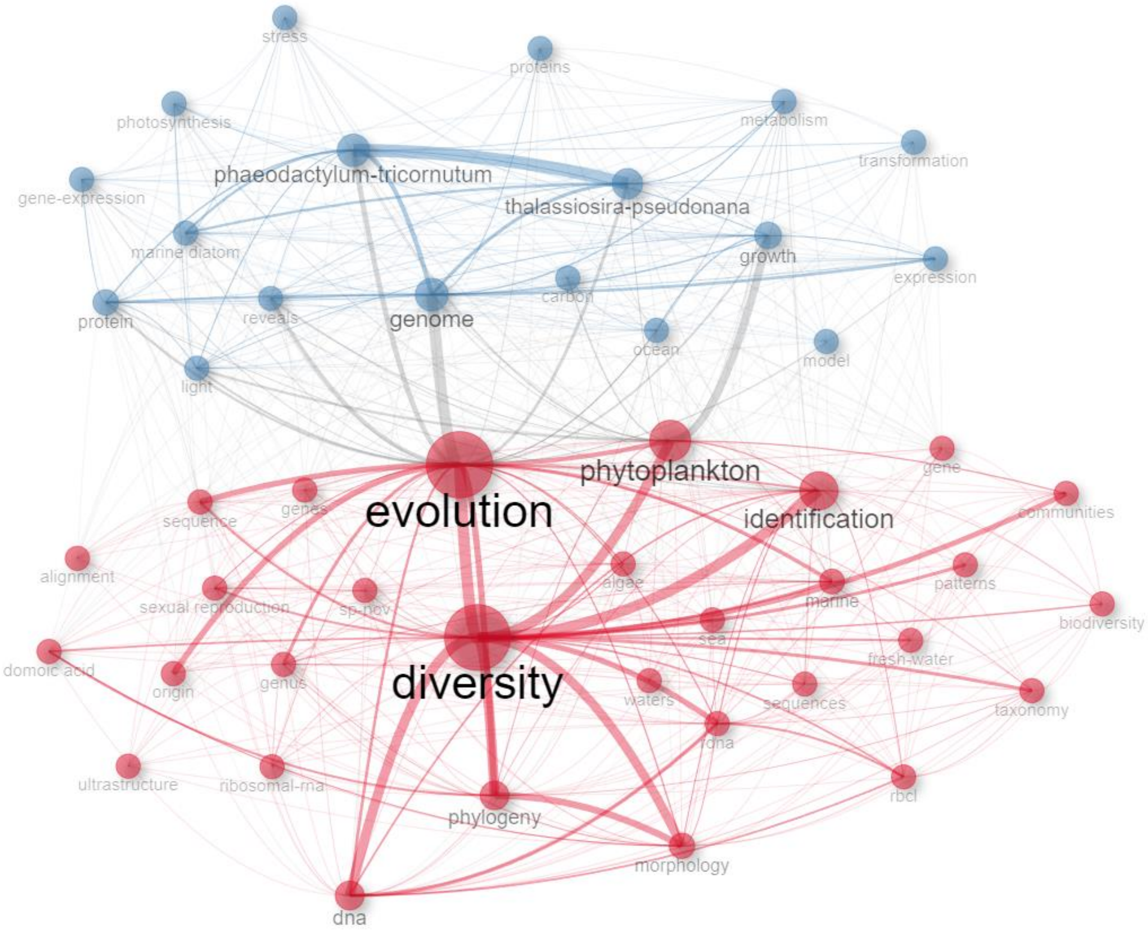


Figure S7. Network based on co-occurence of the most frequent Keywords Plus© (n = 50) in the documents of the *mol-eco-evo* dataset (i.e., subset of molecular diatom research broadly focused on diversity, ecology, and evolution). Contrary to the previous analyses (Figs S5-6), the unit of analysis here were not the documents themselves, but the keywords found within these documents. The nodes represent individual keywords, while their sizes are proportional to their frequency in the dataset. The thickness of the edges reflect frequency of co-occurance of a given keywords. The analysis distinguished two clusters, which we interpreted as (gen)omics (blue) and molecular taxonomy (red).

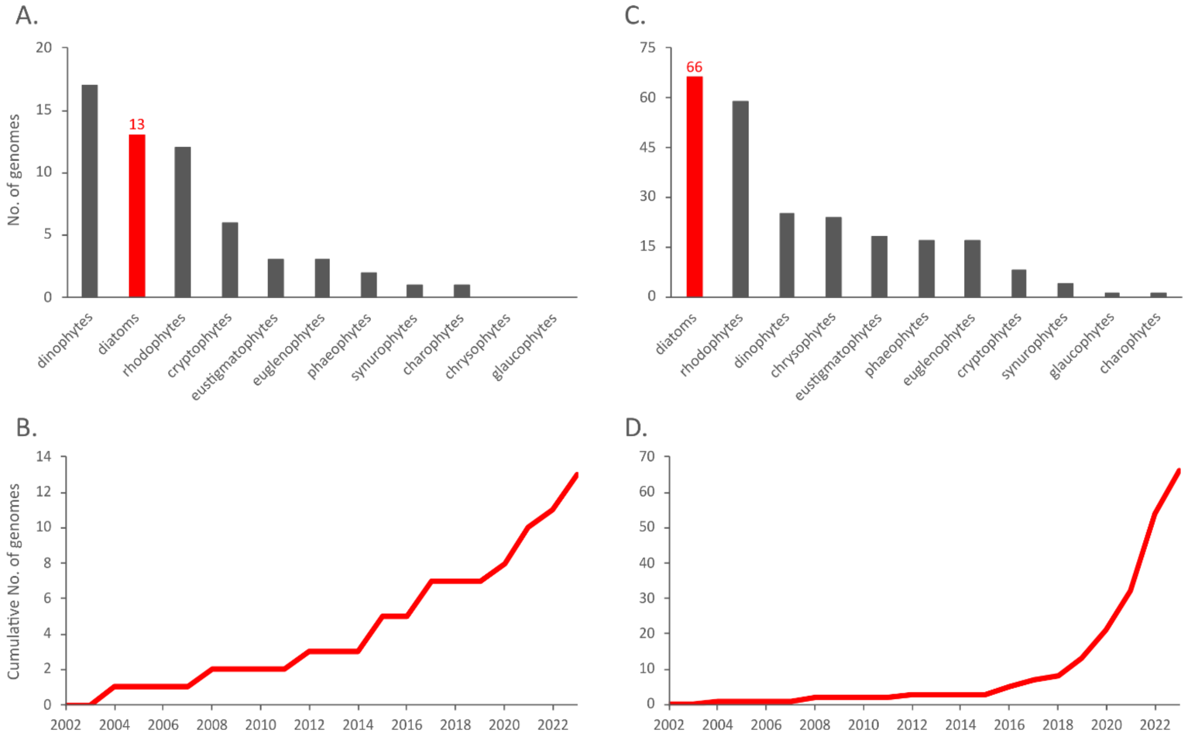


Figure S8. A quantitative comparison of algal genomics. Panels A and C depict comparison of diatoms to other algal groups in number of whole annotated genomes (A), and all nuclear genomes present in the database (meaning also unpublished nuclear genomes, MAGs etc.; C). For the sake of depiction, chlorophytes *s.s.* are not shown (58 and 257 genomes in A and C, respectively). Panels B and D depict a development of the number of genomes over time (whole annotated genomes in B, and all nuclear genomes in D). The figures are based on data from GenBank (https://www.ncbi.nlm.nih.gov/datasets/genome; accessed on the 1st of January 2024).