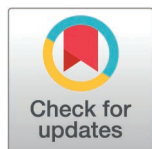


The 100 Diatom Genomes Project

Thomas Mock^{1,2*}, Gust Bilcke^{3,4,5}, Elliott Flaum⁶, Shunan Fu¹, Lilian Hoch⁷, Kevin Moog⁸, Nadine Rijdsdijk^{9,10}, Elizabeth C. Ruck⁸, Ian W. Bishop⁹, Carole Duchene¹⁰, Reuben Gilbertson¹, Amanda Hopes¹, Christopher Johns¹¹, Francesco Manfellotto¹², Wade R. Roberts⁸, Nigel Belshaw¹, Udit Chandra¹³, Peter Chaerle³, Olga Chepurnova³, Daan Deleu³, Sofie D'hondt³, Federica Di Costanzo¹⁴, Omayya Dudin¹⁵, Serena Flori¹⁶, Trupti Gaikwad¹⁶, Agnes Groisillier¹³, Andrea Hall¹, Pengyu Ji¹⁷, Louis J. Lavie-Aydat¹³, William H. Lewis¹⁸, Samuel Menicot¹³, Eveline Pinseel³, Eléonore Pottier³, Krisztina Sarkozi¹, Arianna Smerilli¹⁴, Jan Strauss¹⁹, Sander Thierens^{4,5}, Andrew Toseland¹, Yousef Touhami¹⁶, Robert Utting¹, Michiel Van Bel^{4,5}, Cock van Oosterhout¹, Yue Wu¹³, Feng Yang¹³, Erika Allhusen²⁰, John J. Bolton²¹, Chris Bowler²², Thorsten Brinkhoff²³, Anja Poehlein²⁴, Tim Brovarone^{25,26}, Nansheng Chen^{27,28}, Greg Clark²⁹, Matthew D. Clark³⁰, Dario Copetti²⁹, Zongmei Cui^{27,29}, Beini Deng^{27,31}, Jianbo Jian³², Uwe John^{20,33}, Anne D. Jungblut³⁰, JiHeon Kang³⁴, Jon Bent Kristoffersen²⁵, JunMo Lee^{34,35}, Shuya Liu^{27,28}, David G. Mann³⁶, Linda Medlin^{37,38}, Vincent Moulton³⁹, Jelena Radojicic²⁵, Shinya Sato⁴⁰, Rosa Trobajo⁴¹, Klara Wolf²⁰, Norico Yamada⁴², Naihao Ye⁴³, Libin Zhang^{27,31}, Yunyun Zhuang², Gautam Dey⁴⁴, Valeria Di Dato¹⁴, Katherine Helliwell^{16,45}, Marianne Jaubert¹⁰, Peter G. Kroth¹⁷, Marina Montresor¹², Giovanna Romano¹⁴, Tatiana A. Ryneerson⁹, Jayson Talag²⁹, Klaus U. Valentin²⁰, Flora Vincent⁶, Ross F. Waller¹⁸, Glen Wheeler¹⁶, Andrew J. Alverson⁸, Kerrie Barry⁴⁶, LoriBeth Boston^{46,47}, Angela Falciatore¹⁰, Maria Immacolata Ferrante¹², Jie Guo⁴⁶, Jane Grimwood^{46,47}, Richard Hayes⁴⁶, Andrei Herdean⁷, Jerry Jenkins^{46,47}, Min Kim^{46,47}, Wiebe HCF Kooistra¹², Alan Kuo⁴⁶, Anna Lipzen⁴⁶, Nicole Poulsen¹¹, Jeremy Schmutz^{46,47}, Leila Tirichine^{13,48}, Klaas Vandepoele^{4,5,49}, Frédéric Verret²⁵, Wim Vyverman³, Igor V. Grigoriev^{46,50}



OPEN ACCESS

Citation: Mock T, Bilcke G, Flaum E, Fu S, Hoch L, Moog K, et al. (2026) The 100 Diatom Genomes Project. PLoS Biol 24(9): e3003947. <https://doi.org/10.1371/journal.pbio.3003947>

Published: September 1, 2026

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the [Creative Commons CC0](https://creativecommons.org/publicdomain/zero/1.0/) public domain dedication.

Funding: The work (proposal: <https://doi.org/10.46936/10.25585/60001393>) conducted by the U.S. Department of Energy Joint Genome Institute (<https://ror.org/04xm1d337>), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.TM acknowledges support from the Natural Environment Research Council UK (NE/Z504130/1, NE/W005654/1, NE/R015449/2, and NE/T000848/1). GB, ST, MVB, NR, KV, EPI, EPo, DL, and WV acknowledge funding from Research Foundation Flanders (1228423N, 11L2325N, 1221323N, I000323N,

1 School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, United Kingdom, **2** Key Laboratory of Marine Environment and Ecology, Ministry of Education, Ocean University of China, Qingdao, China, **3** Laboratory of Protistology and Aquatic Ecology, Department of Biology, Ghent University, Ghent, Belgium, **4** Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium, **5** Center for Plant Systems Biology, VIB, Ghent, Belgium, **6** Developmental Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany, **7** Climate Change Cluster, Faculty of Science, University of Technology Sydney, Sydney, New South Wales, Australia, **8** Department of Biological Sciences, University of Arkansas, Fayetteville, Arkansas, United States of America, **9** Graduate School of Oceanography, University of Rhode Island, Narragansett, Rhode Island, United States of America, **10** Laboratoire de Photobiologie et Physiologie des Plastides et des Microalgues, UMR7141, CNRS, Sorbonne Université, Institut de Biologie Physico-Chimique, Paris, France, **11** B CUBE - Center for Molecular Bioengineering, TUD Dresden University of Technology, Dresden, Germany, **12** Integrative Marine Ecology, Stazione Zoologica Anton Dohrn, Napoli, Italy, **13** Department of Biology, Nantes Université, CNRS, US2B, UMR 6286, Nantes, France, **14** Ecosustainable Marine Biotechnology, Stazione Zoologica Anton Dohrn, Napoli, Italy, **15** Department of Biochemistry, Faculty of Sciences, University of Geneva, Geneva, Switzerland, **16** Marine Microbiome Department, Marine Biological Association, Plymouth, United Kingdom, **17** Department of Biology, University of Konstanz, Konstanz, Germany, **18** Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom, **19** Research Data Management Department, Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany, **20** Division Bioscience, Department Chemical Ecology, Alfred-Wegener-Institute for Polar and Marine Research, Bremerhaven, Germany, **21** Department of Biological Sciences, University of Cape Town, Rondebosch, Cape Town, South Africa, **22** Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM, Université PSL, Paris, France, **23** Institute for Chemistry and Biology of the Marine Environment (ICBM), University of Oldenburg, Oldenburg, Germany, **24** Department of Genomic and Applied Microbiology & Göttingen Genomics Laboratory, University of Göttingen, Göttingen, Germany, **25** Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research, Heraklion, Greece, **26** School of Biotechnological Engineering, Catholic University of Lyon, Lyon, France,

G001521N, and 01G01323), UGent (BOF-GOA 01G01323, BOF/STA/202409/019), and the European Research Council (DIADAPT, 101160805), as well as the Belgian Science Policy (Belspo) for supporting the BCCM/DCG diatoms culture collection, and EMBRC Belgium-FWO project GOH3817N for infrastructure funding. NC, ZC, and SL were supported by the National Natural Science Foundation of China (NSFC, No. 42561144237) and e-ASIA (No. DM1414). TM, SF, and YZ acknowledge support by the National Natural Science Foundation of China (42276134). SF acknowledges support by China Scholarship Council (202106330017). GR, VDD, and FDC acknowledges support by Ministero degli Affari Esteri e della Cooperazione Internazionale Italia (PGR05972). EF is supported by the EIPOD-LinC postdoctoral fellowship programme. GD acknowledges the European Commission (KaryodynEVO, 101078291) and the support of the EMBO Young Investigator Programme. GD and EF acknowledge EMBL for core funding. NP was supported by the Deutsche Forschungsgemeinschaft (PO 2256/1-1). CJ was supported by an MSCA Postdoctoral Fellowship (GA101153556). KEH acknowledges support by the Biotechnology and Biological Sciences Research Council UK (BB/W006286/1), European Research Council (101170086) and Natural Environment Research Council UK (NE/R015449/2). GW acknowledges support from Natural Environment Research Council UK (NE/T000848/1). LT acknowledges support from Région Pays de la Loire Connect Talent EpiAlg and the Epicycle ANR project (ANR-19-CE20-00). MIF was supported by the Italian Ministry of University and Research (No. 202295S3WC). JML acknowledges support from the National Research Foundation of Korea (RS-2023-00209930) and from the Korea Environment Industry & Technology Institute, funded by Ministry of Climate, Energy and Environment (RS-2021-KE001788). FV, JR, JBK, and TB acknowledge support by the Hellenic Foundation for Research and Innovation (RADIO/483), the European Union (GHaNA/734708) and the European Regional Development Fund co-financed by Greece and the European Union (NSRF 2014-2020/CMBR/MIS 5002670). TB was funded by the German Research Foundation (DFG) through the Transregional Collaborative Research Centre 'Roseobacter' (TRR 51). AJA, ECR, and WRR were supported by the U.S. National

27 Laboratory of Marine Ecology and Environmental Sciences, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China, 28 Laboratory for Marine Ecology and Environmental Science, Qingdao Marine Science and Technology Center, Qingdao, China, 29 Arizona Genomics Institute, University of Arizona, Tucson, Arizona, United States of America, 30 Department of Sciences, Natural History Museum, London, United Kingdom, 31 College of Marine Science, University of Chinese Academy of Sciences, Beijing, China, 32 Guangdong Provincial Key Laboratory of Marine Biotechnology, Shantou University, Shantou, China, 33 Biodiversity Change, Helmholtz Institute for Functional Marine Biodiversity, Oldenburg, Germany, 34 Department of Oceanography, Kyungpook National University, Daegu, South Korea, 35 Kyungpook Institute of Oceanography, Kyungpook National University, Daegu, South Korea, 36 Science Division, Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom, 37 Algae Save the World, Consulting, Cardigan, United Kingdom, 38 Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, United Kingdom, 39 School of Computing Sciences, University of East Anglia, Norwich, United Kingdom, 40 Faculty of Marine Science and Technology, Fukui Prefectural University, Obama, Fukui, Japan, 41 Marine and Continental Waters Programme, IRTA-Institute for Food and Agricultural Research and Technology, La Ràpita, Spain, 42 Adaptation and Diversity in the Marine Environment, Roscoff Marine Station, French National Centre for Scientific Research, Roscoff, France, 43 National Key Laboratory of Mariculture Biobreeding and Sustainable Goods, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, China, 44 Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, Heidelberg, Germany, 45 Biosciences, Faculty of Health and Life Sciences, University of Exeter, Exeter, United Kingdom, 46 US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, 47 Genome Sequencing Center, HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, United States of America, 48 Institute for Marine and Antarctic Studies (IMAS), Ecology and Biodiversity Centre, University of Tasmania, Hobart, Tasmania, Australia, 49 Center for AI & Computational Biology, VIB, Ghent, Belgium, 50 Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, California, United States of America

* t.mock@uea.ac.uk

One hundred diatom species have been selected for genome and transcriptome sequencing. The 100 Diatom Genomes Project aims to provide a scalable framework for understanding diatom biodiversity, ecology and evolution, and for investigating their use in biotechnology.

Introduction

Diatoms, the most species-rich algal division, are found in all aquatic and some terrestrial environments, and often dominate primary production. They contribute ~20% of annual global carbon fixation and drive the biological carbon pump sequestering organic carbon to the deep sea [1], yet current genomic sampling is heavily biased toward a limited number of model species [2]. Recognizing their fundamental role in sustaining Earth's habitability, the 100 Diatom Genomes Project (100DGP) was established as an international, interdisciplinary project, bringing together 108 researchers from 11 countries to leverage diatom genomics across a broad range of scientific disciplines. Expanding the genome representation across diverse diatom clades, ecological niches and life histories will facilitate high-resolution comparative genomics to trace evolutionary innovations.

Science Foundation (Grant DEB-2331644). TAR acknowledges support from the US National Science Foundation (award 2227425). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Approaches to studying diatoms in the 100DGP

Species selection was based on practical considerations, including strain availability in culture collections and diatom laboratories from around the world, intrinsic growth rates, and the feasibility of minimizing prokaryotic contaminants. Further checkpoints were the quality and quantity of extractable gDNA and total RNA, both of which can differ substantially between species. So far, representative strains of 104 diatom species have passed our suitability assessments and, of those, 20 annotated assemblies are available (Fig 1a; <https://doi.org/10.5281/zenodo.21360381>). Although these 104 diatoms cover all major clades, it is not possible to comprehensively sample the entire diversity of diatoms owing to strain availability and cultivability. To mitigate these constraints, we aim to continue the project beyond 100 diatom genomes. The [100DGP portal](#) provides users with access to all genome assemblies, transcriptomes and annotations. A combination of long-read (PacBio HiFi CCS) and short-read (Illumina NovaSeq X & S4) sequencing has been applied to generate haplotype-resolved genome assemblies and organellar genomes, and a diverse set of transcriptomes (IsoSeq, RNAseq) have been generated for all species for genome annotations. Moreover, for a subset of species, transcriptomes have been generated from cultures exposed to light/dark cycles or raised temperature treatments for future comparative functional studies (Illumina NovaSeq) (Fig 1b; <https://doi.org/10.5281/zenodo.21360381>). All genomes sequenced at the Joint Genome Institute (JGI) were assembled with hifiasm [5], polished with Racon [6], and annotated using the JGI annotation pipeline [7]. To provide user-friendly interfaces for comparative genomics, multiple bioinformatics analyses were performed and integrated through the PhycoCosm [8] (JGI, Berkeley, USA) and PLAZA platforms (VIB, Ghent, Belgium) (Fig 1c–1f).

To bridge the gap between genetic variants and key underpinning traits, we are complementing our omics analyses with light and expansion microscopy (ExM) and high-throughput phenotyping (phenomics) (Fig 2). As a technique that enables researchers to physically enlarge diatoms, ExM will facilitate the development of molecular markers to resolve the subcellular localization of proteins with nanometer-scale precision, including those encoded by previously uncharacterized or lineage-specific genes. Our high-throughput phenotyping platform should reveal if and how selected proteins have an impact on measurable traits (e.g., carbon acquisition) and their phenotypic plasticity (Fig 2b). The latter will be quantified using the environmentally standardized plasticity index (*ESPI), a recently developed metric that integrates trait responses across controlled environmental gradients to capture the breadth and magnitude of species-specific plastic responses [9]. By applying the *ESPI, we will compare trait plasticity between species and link differences in their plastic response to ecological niches, global distribution patterns, and therefore adaptive potential. Thus, this integrated framework will lay the foundation for mechanistic trait ontology, which is essential to understand both the evolutionary history of diatoms and the adaptability of important traits.

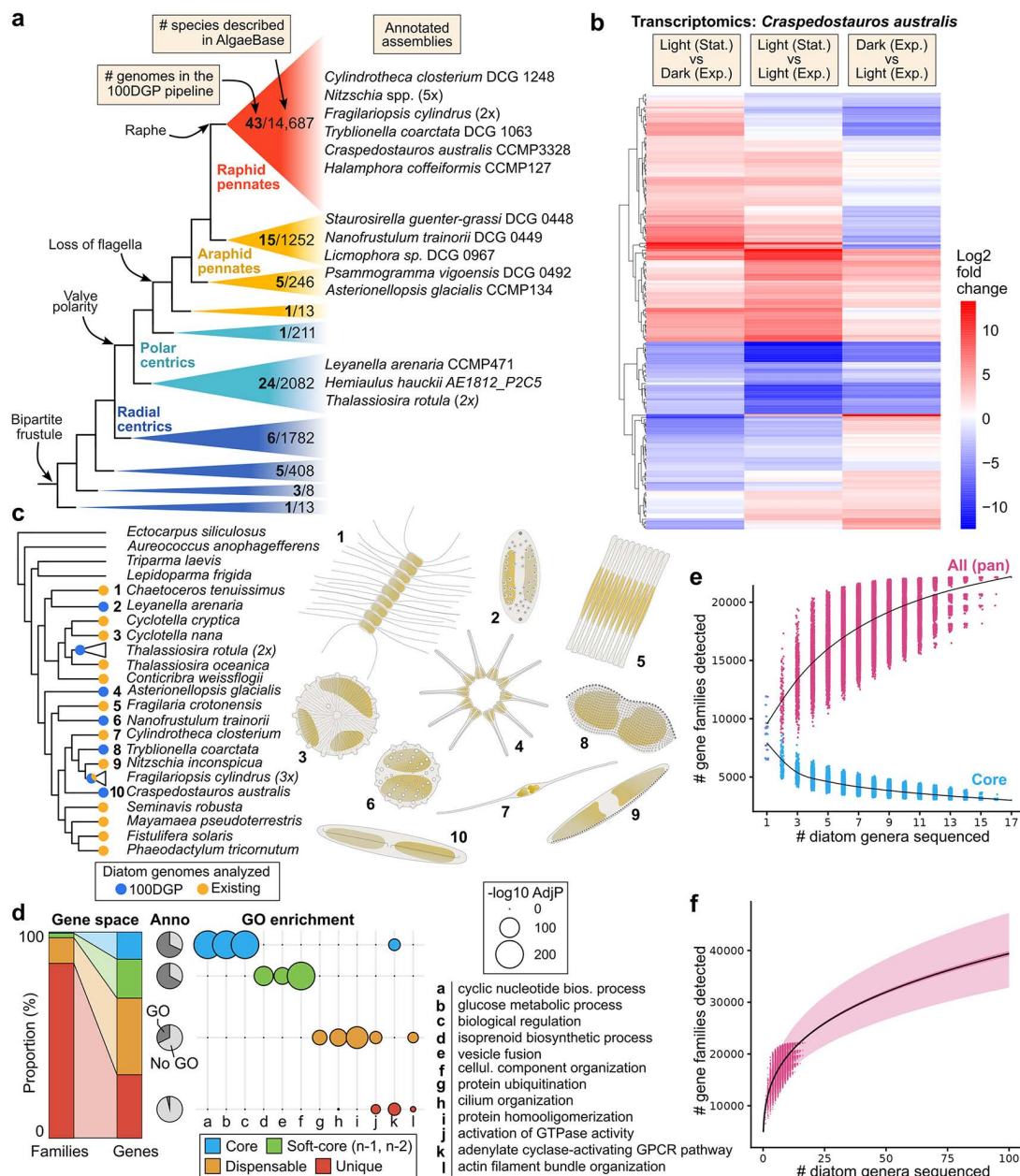


Fig 1. Species selection and gene space diversity of the 100 Diatoms Genome Project. **a.** Simplified drawing of the 10 clades of the diatom phylogeny following [2]. Within each clade, the number of species in the pipeline of the 100 Diatom Genomes Project (100DGP) is indicated, as well as the number of species currently described in *AlgaeBase* (searched on 3 May 2026). The names of species for which both assembly and annotation of genomes have been completed are shown on the right. **b.** Example of transcriptomic data generated as part of the 100DGP, illustrated for *Craspedostauros australis*. The heatmap shows the log₂ fold changes of 193 diel-responsive genes identified for cells grown under a 14 h light:10 h dark photoperiod (adjusted $P < 0.001$; $\log_2 FC > 2$ or $\log_2 FC < -2$). Cells were harvested for RNA isolation at the midpoint of the light phase and the midpoint of the dark phase on day 4, when cultures were growing exponentially ('Exp.'), and at the midpoint of the light phase on day 12, when cultures had reached stationary-phase and formed a biofilm ('Stat.'). Gene expression patterns reveal transcriptional responses associated with the diel cycle as well as changes accompanying the transition from exponential growth to stationary-phase biofilm development. **c.** Cladogram of selected diatom and outgroup species used to assess gene space evolution, based on the phylogenetic relationships from [2]. In total, 22 diatom genomes covering 17 genera were selected for a preliminary comparative genomic analysis, combining the 9 first annotated genomes from the 100DGP and 13 publicly available assemblies. Drawings (not to scale) of selected species illustrate the morphological and ultrastructural diversity of diatoms. **d.** Bar plots showing the proportion of gene families with a distinct distribution over diatom genomes: 'core', appearing in all diatom genomes; 'soft-core', appearing in all but one or two

diatom genomes; 'dispensable', appearing in 2–19 diatom genomes; and 'unique', appearing in a single diatom genome. For each distribution class, pie charts show the proportion that is annotated by at least one gene ontology (GO) term, whereas bubble plots visualize the enrichment significance of the top 3 most enriched GO terms, deduplicated with rrvgo [3]. AdjP: false discovery rate-adjusted p-value. **e.** Pan-genome rarefaction curve [4] showing the number of gene families identified when the number of sequenced diatom genera increases. Both the total number of gene families detected (pink, pan) and the number of families detected in all genera (blue, core) are visualized. **f.** Heap's law (black line) predicts the expected total number of gene families discovered when an increasing number of diatom genera is sampled. The light pink shade depicts the prediction interval, darker pink highlights the confidence interval.

<https://doi.org/10.1371/journal.pbio.3003947.g001>

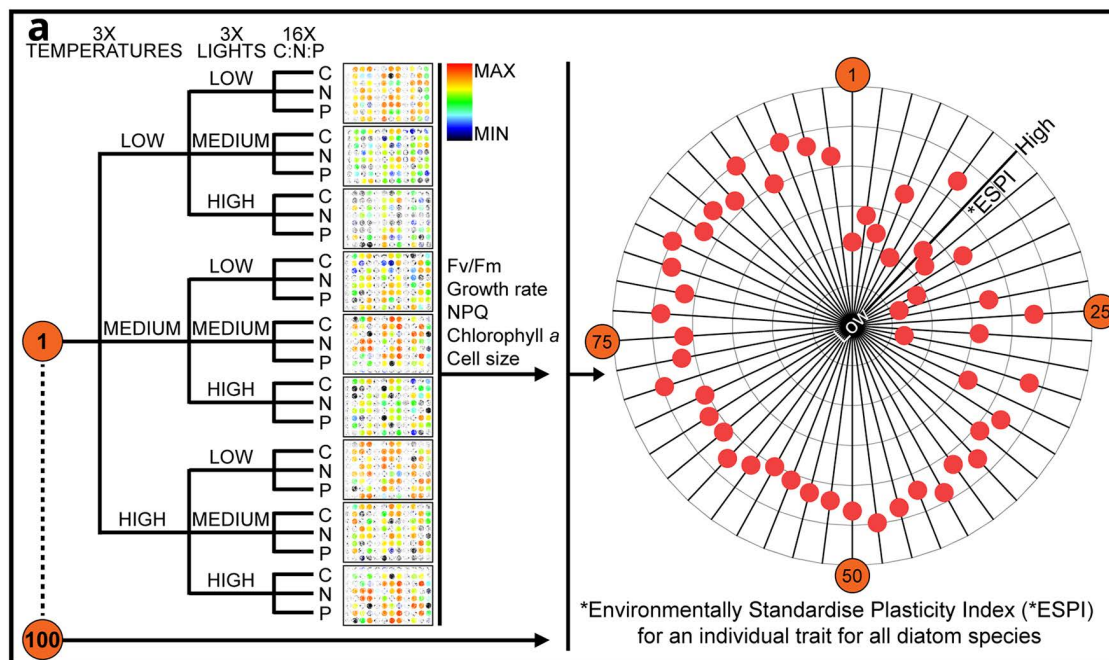
Resources and preliminary insights for advancing integrative diatom research

In line with Findable, Accessible, Interoperable and Reusable (FAIR) principles of scientific data management, the 100DGP will publish laboratory methods including standard operating procedures, code, and a diverse set of resources (e.g., genomes, transcriptomes, images, and trait data), together forming an integrated dataset that can be leveraged for future functional studies of diatoms. We encourage researchers to make use of these resources, including the deposited diatom strains at the [BCCM/DCG diatoms collection](#) (Ghent, Belgium). Some strains are also available from [NCMA](#) (Bigelow, East Boothbay, Maine, USA), [RCC](#) (Roscoff, France) and [CCAP](#) (Oban, UK). Data integration from genomes, transcriptomes, nanoscale imaging and phenomics is unprecedented for any protist group and will allow us to place comparative evolutionary genomics into the context of cell biology, metabolism and ecology for questions that go beyond the role of individual genes. The first genome assemblies are already publicly available ([100DGP portal](#)), revealing a range of genome sizes (75–1,033 Mbp) and >3-fold variation in gene counts. Future comparative analyses will be released through PhycoCosm [8], DiatOmicBase [10], and [PLAZA Diatoms](#): versatile gene-centered platforms for integrative diatom omics research.

External researchers can use this resource to link the outcome of comparative evolutionary genomics (e.g., genomic loci under selection) with nanometer-scale detection of biomolecules and how they shape key traits such as carbon acquisition, nutrient utilization, morphotype diversification, and life-cycle regulation. Beyond fundamental research, our integrative datasets can be exploited for linking genetic variants with the in-vivo function of their encoded proteins and other bioactive compounds for advancing technological applications, such as drug discovery or bioprocess optimization [11].

Preliminary insights based on a comparative analysis of 22 new and existing diatom genomes revealed that only <3% of gene families and 32% of genes are conserved amongst all diatoms (core, soft-core) (Fig 1d). Gene families were identified by performing an all-versus-all homology search between diatom and outgroup proteomes using DIAMOND with subsequent clustering into gene families using the tribeMCL algorithm. Since the rise of diatoms more than 250 mya, a high level of genetic diversity has evolved through a combination of divergent evolution in existing gene families and the appearance of novel genes, either through de-novo gene birth or horizontal gene transfer [12]. Although most gene ontology terms enriched in the diatom shared pan-genome represent genes involved in primary metabolism, the enrichment of cyclic nucleotide-mediated signaling in the diatom core genome supports previous evidence that these secondary messengers drive a multitude of essential abiotic and biotic processes in diatoms. Dispensable and unique genes, on the other hand, were enriched in receptor signaling, flagella organization during reproduction, cell division, DNA recombination, and nuclear organization, which suggest that these processes underpin speciation. Although still preliminary, these results suggest that adaptation may have led to the emergence of novel genes and the exceptional diversity of diatoms.

Based on the current number of genomes, it appears the diatom pan-genome is open (Fig 1e), and Heap's law [13] predicts that the number of gene families, including their novel genes and variants, will double to ~40,000 by the project end (Fig 1f). To address their potential role in diatom cell biology, particularly if identified as part of known metabolic pathways such as photosynthesis, we will develop molecular markers to reveal the subcellular localization of their encoded proteins using ExM [14]. Our high-throughput phenotyping platform will show if and how those genetic variants and their encoded proteins affect measurable traits, such as light harvesting, carbon acquisition, elemental stoichiometry, or cell



Trait Ontology for Diatoms

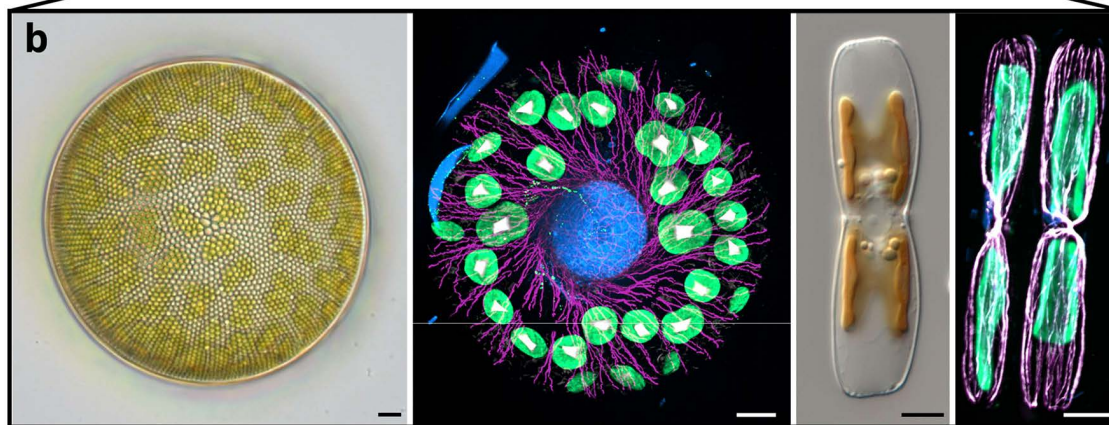


Fig 2. Phenomics, imaging and cellular physiology. a. High-throughput phenotyping pipeline to quantify key traits (maximum PSII quantum yield (Fv/Fm), growth rate, cell size, chlorophyll-a content, and non-photochemical quenching (NPQ)), under different environmental conditions (3x temperatures, 3x light intensities, 16x C:N:P ratios). These data will be used to quantify phenotypic plasticity using an Environmentally Standardised Plasticity Index (*ESPI) representing the breadth of the phenotypic responses of each species to changing environmental conditions. This metric is designed to facilitate cross-species comparisons of phenotype plasticity, allowing for greater understanding of phenome response to climate variation. Greatest variance in response per trait is normalized using the Euclidean distance between the environments in which these responses are found, to quantify plasticity per unit of environmental change. In the radial *ESPI plot, each red dot represents the *ESPI value for a given trait for each of the 100 diatom species; larger overall *ESPI values indicate greater phenotypic plasticity. The radial plot shows low *ESPI in the center and high towards the exterior. Information from (a) and (b) will be merged to develop a trait ontology for diatoms. b. Cell morphology of *Coscinodiscus granii* (left) and *Craspedostauros australis* (right) using light and expansion microscopy (ExM) to demonstrate the nanoscale distribution of molecular complexes: tubulin (magenta), photosystem II (green), nucleus and silica cell wall (blue), NHS ester (white). Black scale bars are 5 μm . White scale bars are 20 μm (4.65 μm biological scale) and 5 μm (1.37 μm biological scale), respectively.

<https://doi.org/10.1371/journal.pbio.3003947.g002>

division. Early findings indicate that diatoms exhibit a wide range of phenotypic responses, likely driven by their genetic diversity and cellular structural complexity. Altogether, we are constructing a common ontology of traits, critical for developing an integrated approach based on evolutionary genomics, suitable for building predictive models from cell biology to ecosystem science.

Conclusions

The 100DGP aims to generate diverse biological resources for a key group of primary producers. Our project also aims to provide a test case for protist research at scale, which, for many other protist groups, is still in its infancy. To grow beyond 100 diatom genomes, we are integrating externally sequenced diatom genomes and transcriptomes from other projects (e.g., from the *Chaetoceros* genus sequencing project at JGI). Our preliminary results, based on all resources, provide evidence that integrated research at scale is feasible for diatoms and promise to be a potential treasure trove for discovering novel biology, including unique adaptations and interactions (e.g., endosymbionts). We hope that this work will improve our understanding of global species diversity, evolution, and ecosystem functioning, as well as providing novel genetic variants for advancing biotechnology via genetic engineering and synthetic biology.

Acknowledgments

TM acknowledges the School of Environmental Sciences, University of East Anglia, Norwich (UK), and the College of Environmental Science and Engineering, Ocean University of China for providing time to coordinate this large-scale project. NP acknowledges the CMCB light microscopy facility, a Core Facility of the CMCB Technology Platform at TU Dresden (Germany) for light microscopy support.

Author contributions

Conceptualization: Thomas Mock, Gust Bilcke, Gautam Dey, Andrew J. Alverson, Angela Falciatore, Maria Immacolata Ferrante, Andrei Herdean, Wiebe HCF Kooistra, Nicole Poulsen, Klaas Vandepoele, Wim Vyverman, Igor V. Grigoriev.

Data curation: Thomas Mock, Gust Bilcke, Elliott Flaum, Lilian Hoch, Amanda Hopes, Jan Strauss, Andrew Toseland, Anja Poehlein, Angela Falciatore, Maria Immacolata Ferrante, Richard Hayes, Andrei Herdean, Alan Kuo.

Formal analysis: Gust Bilcke, Elliott Flaum, Lilian Hoch, Christopher Johns, Serena Flori, Andrew Toseland, Jianbo Jian, Marina Montresor, Jie Guo, Richard Hayes, Jerry Jenkins, Min Kim, Alan Kuo, Anna Lipzen, Nicole Poulsen.

Funding acquisition: Thomas Mock, Elizabeth C. Ruck, Wade R. Roberts, Omayya Dudin, Eveline Pinseel, Cock van Oosterhout, Thorsten Brinkhoff, Matthew D. Clark, Anne D. Jungblut, Linda Medlin, Vincent Moulton, Naihao Ye, Yunyun Zhuang, Gautam Dey, Katherine Helliwell, Peter G. Kroth, Marina Montresor, Giovanna Romano, Tatiana A. Ryneerson, Klaus U. Valentin, Flora Vincent, Ross F. Waller, Glen Wheeler, Andrew J. Alverson, Angela Falciatore, Maria Immacolata Ferrante, Wiebe HCF Kooistra, Nicole Poulsen, Leïla Tirichine, Klaas Vandepoele, Frédéric Verret, Wim Vyverman.

Investigation: Thomas Mock, Gust Bilcke, Elliott Flaum, Shunan Fu, Kevin Moog, Nadine Rijdsdijk, Francesco Manfellotto, Wade R. Roberts, Udit Chandra, Daan Deleu, Federica Di Costanzo, Serena Flori, Trupti Gaikwad, Agnes Groisillier, Andrea Hall, Louis J. Lavie-Aydat, William H. Lewis, Samuel Menicot, Eveline Pinseel, Eléonore Pottier, Arianna Smerilli, Yousef Touhami, Yue Wu, Feng Yang, Tim Brovarone, Jon Bent Kristoffersen, Shuya Liu, Jelena Radojicic, Giovanna Romano, Tatiana A. Ryneerson, Jayson Talag, LoriBeth Boston, Jane Grimwood, Richard Hayes, Alan Kuo, Nicole Poulsen, Leïla Tirichine, Frédéric Verret, Wim Vyverman.

Methodology: Thomas Mock, Shunan Fu, Lilian Hoch, Nadine Rijdsdijk, Elizabeth C. Ruck, Ian W. Bishop, Carole Duchene, Reuben Gilbertson, Amanda Hopes, Wade R. Roberts, Nigel Belshaw, Daan Deleu, Sofie D'hondt, Omayya

Dudin, Trupti Gaikwad, Andrea Hall, Pengyu Ji, Louis J. Lavier-Aydat, Samuel Menicot, Eveline Pinseel, Eléonore Pottier, Krisztina Sarkozi, Jan Strauss, Andrew Toseland, Yousef Touhami, Anja Poehlein, Tim Brovarone, Matthew D. Clark, Zongmei Cui, Jianbo Jian, Uwe John, Jon Bent Kristoffersen, Jelena Radojicic, Klara Wolf, Gautam Dey, Marianne Jaubert, Peter G. Kroth, Tatiana A. Ryneerson, Jayson Talag, Flora Vincent, Angela Falcitore, Jane Grimwood, Andrei Herdean, Jerry Jenkins, Min Kim, Wiebe HCF Kooistra, Nicole Poulsen, Jeremy Schmutz, Leïla Tirichine, Klaas Vandepoele, Frédéric Verret, Wim Vyverman, Igor V. Grigoriev.

Project administration: Thomas Mock, Gust Bilcke, Peter Chaerle, Andrea Hall, Erika Allhusen, Thorsten Brinkhoff, Gautam Dey, Katherine Helliwell, Klaus U. Valentin, Flora Vincent, Kerrie Barry, Jane Grimwood, Jeremy Schmutz, Frédéric Verret, Wim Vyverman, Igor V. Grigoriev.

Resources: Thomas Mock, Nadine Rijdsdijk, Carole Duchene, Amanda Hopes, Nigel Belshaw, Peter Chaerle, Olga Chepuranova, Jan Strauss, Sander Thierens, Robert Utting, Michiel Van Bel, Erika Allhusen, John J. Bolton, Thorsten Brinkhoff, Nansheng Chen, Greg Clark, Uwe John, Shuya Liu, David G. Mann, Shinya Sato, Rosa Trobajo, Klara Wolf, Norico Yamada, Naihao Ye, Yunyun Zhuang, Valeria Di Dato, Marianne Jaubert, Peter G. Kroth, Marina Montresor, Giovanna Romano, Jayson Talag, LoriBeth Boston, Maria Immacolata Ferrante, Jane Grimwood, Nicole Poulsen, Jeremy Schmutz, Leïla Tirichine, Klaas Vandepoele, Frédéric Verret, Wim Vyverman, Igor V. Grigoriev.

Software: Sander Thierens, Michiel Van Bel, Jie Guo.

Supervision: Thomas Mock, Elizabeth C. Ruck, Amanda Hopes, Peter Chaerle, Eveline Pinseel, Krisztina Sarkozi, Thorsten Brinkhoff, Dario Copetti, Jianbo Jian, Vincent Moulton, Naihao Ye, Gautam Dey, Valeria Di Dato, Katherine Helliwell, Marianne Jaubert, Peter G. Kroth, Giovanna Romano, Tatiana A. Ryneerson, Jayson Talag, Flora Vincent, Ross F. Waller, Glen Wheeler, Angela Falcitore, Maria Immacolata Ferrante, Jane Grimwood, Andrei Herdean, Jeremy Schmutz, Leïla Tirichine, Klaas Vandepoele, Frédéric Verret, Wim Vyverman, Igor V. Grigoriev.

Validation: Thomas Mock, Elliott Flaum, Kevin Moog, Wade R. Roberts, Anja Poehlein.

Visualization: Thomas Mock, Elliott Flaum, Lilian Hoch, Kevin Moog, Amanda Hopes, Christopher Johns, Serena Flori, Trupti Gaikwad, Yousef Touhami, Nicole Poulsen.

Writing – original draft: Thomas Mock, Andrei Herdean.

Writing – review & editing: Thomas Mock, Elliott Flaum, Lilian Hoch, Nadine Rijdsdijk, Amanda Hopes, Peter Chaerle, Daan Deleu, Eveline Pinseel, Eléonore Pottier, Krisztina Sarkozi, Jan Strauss, Cock van Oosterhout, Chris Bowler, Nansheng Chen, Matthew D. Clark, Zongmei Cui, Beini Deng, Jianbo Jian, JiHeon Kang, JunMo Lee, Linda Medlin, Vincent Moulton, Norico Yamada, Naihao Ye, Libin Zhang, Yunyun Zhuang, Gautam Dey, Valeria Di Dato, Katherine Helliwell, Peter G. Kroth, Marina Montresor, Giovanna Romano, Tatiana A. Ryneerson, Flora Vincent, Ross F. Waller, Glen Wheeler, Andrew J. Alverson, Angela Falcitore, Maria Immacolata Ferrante, Andrei Herdean, Min Kim, Wiebe HCF Kooistra, Nicole Poulsen, Leïla Tirichine, Klaas Vandepoele, Frédéric Verret, Wim Vyverman, Igor V. Grigoriev.

References

1. Field C, Behrenfeld M, Randerson J, Falkowski P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*. 1998;281(5374):237–40. <https://doi.org/10.1126/science.281.5374.237> PMID: 9657713
2. Alverson AJ, Roberts WR, Ruck EC, Nakov T, Ashworth MP, Brylka K, et al. Phylogenomics reveals the slow-burning fuse of diatom evolution. *Proc Natl Acad Sci U S A*. 2025;122(22):e2500153122. <https://doi.org/10.1073/pnas.2500153122> PMID: 40440071
3. Sayols S. rrvgo: a bioconductor package for interpreting lists of gene ontology terms. *MicroPubl Biol*. 2023;2023:10.17912/micropub.biology.000811. <https://doi.org/10.17912/micropub.biology.000811> PMID: 37151216
4. Zhao Y, Jia X, Yang J, Ling Y, Zhang Z, Yu J, et al. PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics*. 2014;30(9):1297–9. <https://doi.org/10.1093/bioinformatics/btu017> PMID: 24420766
5. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;18(2):170–5. <https://doi.org/10.1038/s41592-020-01056-5> PMID: 33526886

6. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 2017;27(5):737–46. <https://doi.org/10.1101/gr.214270.116> PMID: [28100585](#)
7. Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otilar R, et al. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* 2014;42(Database issue):D699–704. <https://doi.org/10.1093/nar/gkt1183> PMID: [24297253](#)
8. Grigoriev IV, Hayes RD, Calhoun S, Kamel B, Wang A, Ahrendt S, et al. PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Res.* 2021;49(D1):D1004–11. <https://doi.org/10.1093/nar/gkaa898> PMID: [33104790](#)
9. Hoch L, Herdean A, Woodcock S, Songsomboon K, Osborne B, Ralph PJ. New method for quantification of phenotypic plasticity reveals how plasticity changes over time in the Diatom *Thalassiosira weissflogii*. *Ecol Evol.* 2026;16(2):e73072. <https://doi.org/10.1002/ece3.73072> PMID: [41716591](#)
10. Villar E, Zweig N, Vincens P, Cruz de Carvalho H, Duchene C, Liu S, et al. DiatOmicBase: a versatile gene-centered platform for mining functional omics data in diatom research. *Plant J.* 2025;121(6):e70061. <https://doi.org/10.1111/tpj.70061> PMID: [40089834](#)
11. Dong Z, Xiang W, Jiang W, Guo T. Expansion omics: from expansion microscopy to spatial omics. *Mol Syst Biol.* 2026;22(2):165–78. <https://doi.org/10.1038/s44320-025-00171-9> PMID: [41326779](#)
12. Vancaester E, Depuydt T, Osuna-Cruz CM, Vandepoele K. Comprehensive and functional analysis of horizontal gene transfer events in Diatoms. *Mol Biol Evol.* 2020;37(11):3243–57. <https://doi.org/10.1093/molbev/msaa182> PMID: [32918458](#)
13. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol.* 2008;11(5):472–7. <https://doi.org/10.1016/j.mib.2008.09.006> PMID: [19086349](#)
14. Flori S, Mikus F, Flaum E, Moog K, Guessoum S, Beavis T, et al. Diatom ultrastructural diversity across controlled and natural environments. *Curr Biol.* 2025;35(23):5709–5720.e4. <https://doi.org/10.1016/j.cub.2025.10.024> PMID: [41175869](#)