Insulated Outlier Chromosomes Enable Metabolic Innovation in Minimal Eukaryotic Algae

Martha Valiadi[1,2,*], Keith Harrison[1,3,*], Yann Loe-Mie[4], Bryony Williams[3], Dyan Ankrett[3], Nicholas Smirnoff[3] & Adam Monier[1,3]

[1]Living Systems Institute, University of Exeter, Exeter, UK; [2]Foundation for Research and Technology Hellas, Institute of Molecular Biology and Biotechnology, Heraklion, Crete, Greece; [3]Biosciences, Faculty of Health and Life Sciences, University of Exeter, Exeter, UK; [4]Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France.

*Equal contributions

Correspondence: a.monier@exeter.ac.uk

Abstract

How do organisms with compact genomes maintain evolutionary adaptability? Mamiellales, marine picoeukaryotes including *Ostreococcus tauri*, the smallest known free-living eukaryote, achieve ecological success despite severe genomic constraints. Using chromosome conformation capture (Hi-C) to investigate three-dimensional genome organisation in *O. tauri*, we reveal that two outlier chromosomes with low-GC regions are spatially insulated from the core genome, forming compartments akin to topologically associated domains. These compartments exhibit hypomethylation, altered nucleosome dynamics, and enrichment in transposable elements, introns, and highly expressed genes. They harbour polyketide synthase genes undergoing active structural diversification, with species-specific metabolite profiles including putative polyketides. This genomic architecture enables evolutionary experimentation while protecting essential functions, a strategy we demonstrate is convergently evolved in the distantly related picoeukaryote *Pelagomonas*. Our findings reveal that spatial compartmentalisation of adaptive processes represents a fundamental mechanism for maintaining evolutionary flexibility under genomic streamlining.

Main

The marine green alga *Ostreococcus* is the smallest known free-living eukaryote (~0.8 µm), with a single plastid and mitochondrion[1,2]. This minimalistic cellular organisation presents a simplified eukaryotic model for studying fundamental processes such as photosynthesis, circadian rhythms, and environmental responses[3–5]. *Ostreococcus* belongs to the Mamiellales (class Mamiellophyceae), which includes other picoalgae like *Bathycoccus*, *Mantoniella*, and *Micromonas*[6]. Mamiellales thrive globally across diverse marine environments, from tropical to polar waters[7–9], supported by ecotype differentiation and intra-order genetic diversity[10–12].

Yet this ecological success presents an evolutionary paradox: Mamiellales possess some of the most compact eukaryotic genomes known. *Ostreococcus tauri* has a 12.9 Mb genome encoding ~8,000 genes[13], much smaller and denser than other green algae such as *Chlamydomonas reinhardtii* at 111.1 Mb with 17,742 genes[14]. Their small, haploid, gene-dense genomes[15–21] may confer efficiency under resource limitation[21,22], but such genomic streamlining should theoretically constrain evolutionary flexibility by reducing genetic redundancy and mutational targets[13].

How do organisms with extremely compact genomes maintain the evolutionary adaptability necessary for ecological success? A hallmark feature of Mamiellales genomes may hold the answer: two outlier chromosomes, the so-called big and small outlier chromosomes (BOC and SOC), each containing regions of markedly lower GC content (BOC1 and SOC1)[23]. These atypical regions are enriched in introns, transposable elements (TEs), and highly expressed genes[24–26] and vary significantly among *Ostreococcus* strains[27]. These polymorphisms are associated with phenotypic differences in growth and viral susceptibility[27,28]. In *O. tauri*, the BOC includes a mating-type locus, and genes linked to chromatin conformation[24,27,29], while the SOC, highly variable due to duplication and translocation, is involved in antiviral defence[28,30]. Whether these outlier chromosomes represent an evolutionary solution to maintaining adaptability under extreme genomic constraints remains unknown.

We hypothesised that BOC1 and SOC1 act as spatially insulated hotspots for genomic innovation, enabling adaptation without compromising genome compactness. Their TE and intron richness may promote rearrangement and splicing diversity[31,32], while low GC content may affect chromatin structure and gene regulation[33–35]. Elevated gene expression in these regions suggests a distinct regulatory environment[24,25,36]. To test this, we employed a multi-omics approach–Hi-C, transcriptomics, genome-wide methylation analysis, and metabolomics–to examine the structure, regulation, and functional output of the outlier chromosomes. We provide evidence that picoalgae compartmentalise adaptive functions to resolve the tension between genome streamlining and evolutionary responsiveness.

Results

Structural conservation and functional specialisation of outlier regions across Mamiellales

If BOC1 and SOC1 function as specialised compartments for genomic innovation, they should be conserved across the Mamiellales lineage, indicating functional importance, and enriched in genes involved in genomic rearrangements and regulation. Despite their highly compact genomes (Supplementary Tables 1, 2), all examined Mamiellales genomes (*Ostreococcus* sp. RCC809, *O. tauri*, *O. lucimarinus*, *Micromonas commoda*,

Micromonas pusilla, and *Bathycoccus prasinos*[13,16,17,24]) contain distinct low-GC outlier regions. Even in *Mantoniella tinhauana*, where their presence had been uncertain[37], we identified two such regions with ~15% lower GC content than the surrounding genome (Supplementary Fig. 1; Supplementary Table 2). Although BOC1 and SOC1 regions vary considerably in size across lineages (Fig. 1a), their consistent presence suggests they are an ancient, conserved feature of Mamiellales. Comparative analysis across six Mamiellales species revealed that BOC1 regions maintain greater evolutionary conservation than SOC1. While SOC1 regions showed limited synteny and fewer orthologues across genera (Supplementary Fig. 2), BOC1 regions shared more orthologues, though not strict synteny (Fig. 1a).

The functional specialisation of these regions is evident in their gene content. Gene Ontology enrichment analyses of *O. tauri* BOC1 revealed a striking concentration of genes involved in DNA transposition, transcription, and integration—hallmarks of genomic plasticity and rearrangement activity (Fig. 1b; Extended Data 1). The region also harbours genes for chromatin remodelling, including mini-chromosome maintenance (MCM) complex components and sister chromatid cohesion protein 2 (SCC2). In contrast, the SOC1 shows enrichment for RNA polymerase activity, membrane components, and methylation machinery (Extended Data 1), suggesting distinct but complementary regulatory function.

Transposable elements and introns cluster in outlier genomic regions across Mamiellales genomes

The enrichment of transposition and rearrangement genes in outlier regions suggests these compartments actively drive genomic variation. We conducted a survey of introns and TEs across Mamiellales genomes, confirming previous findings[15,37,38] and revealing a distinct, non-random distribution of these features. In *O. tauri*, BOC1 and SOC1 contain 14.5x and 17.4x more TEs, respectively, than the standard genome (Kolmogorov–Smirnov test, $p < 0.001$; Fig. 2a, b; Supplementary Fig. 3; Extended Data 2). For instance, Ty1/Copia retrotransposon density increases from just 2.04 elements per Mb in the standard genome to 37.4 and 167 elements per Mb in BOC1 and SOC1, respectively. This massive enrichment is not unique to *O. tauri*, with similar patterns occur across most Mamiellales species (Extended Data 2).

TEs often neighbour actively transcribed genes. In *O. tauri*, transcriptome analysis revealed associations between TEs and transcriptionally active genes in BOC1 and SOC1 (Fig. 2c-f; Extended Data 2). For example, a ~5 kb Ty1/Copia retrotransposon sits adjacent to a hexosyltransferase gene (Fig. 2c), while a ~1 kb Tc1/Mariner DNA transposon is near a transcription factor IIA gene (Fig. 2d). Mutator-like elements (MULEs) frequently associate with giant modular polyketide synthase (PKS) genes, sometimes even inserting themselves upstream and downstream of coding segments (Fig. 2e, f).

Intron enrichment provides complementary evidence for enhanced rearrangement potential, with BOC1 regions showing consistently higher intron density across taxa, for instance four-fold higher than the standard genome in *O. tauri* (Extended Data 2). This intron enrichment likely facilitates alternative splicing and exon shuffling, providing additional routes to functional diversification in these compact genomes.

Genes on outlier regions are more highly expressed and frequently spliced

We next asked whether genes in outlier regions are expressed differently. Analysing transcriptomic data from *O. tauri* and other Mamiellales species, we found genes in BOC1 regions show consistently higher expression compared to the standard genome across all species. In *O. tauri*, outlier genes show an overall 3.8-fold higher expression (Kruskal-Wallis, $p<0.001$; Fig. 3a, e). Similar trends are observed in *M. commoda*, *M. pusilla*, and *B. prasinos* (2.4, 3.3, and 1.6-fold, respectively; $p<0.001$) but not all species had higher expression in SOC1 (Extended Data 2). This transcriptional overactivity correlates with the low GC content characteristic of these outlier regions. In *O. tauri*, *M. commoda*, and *M. pusilla*, lower GC content predicts higher expression (Spearman's rank, $\rho=-0.37$, -0.29, and -0.49, $p<0.001$; Fig. 3i, j, k); *B. prasinos* is the only species where the trend is reversed ($\rho=0.22$, $p<0.001$; Fig. 3l).

Beyond high gene expression, outlier regions drive transcriptional diversity through alternative splicing. Despite containing fewer total genes than the standard genome, BOC1 regions in *O. tauri* generate 38% (1,216 of 3,215) of all splice variants. This disproportionate contribution suggests these regions maximise functional output from limited genetic space. *O. lucimarinus*, *B. prasinos* and *M. commoda* all show similar splice variant enrichment (20%; Extended Data 2), confirming this as a conserved feature of outlier region regulation.

Three-dimensional genome architecture reveals insulation of outlier regions

How are TEs concentrated and constrained within Mamiellales outlier regions, while avoiding the essential core genome? To address this question, we used Hi-C to examine the 3D organisation of outlier regions. Hi-C maps (330 million contacts at 10 kb resolution; Fig. 4a; Supplementary Fig. 4) showed a simple architecture dominated by intrachromosomal interactions (77% of all contacts; Extended Data 3). BOC and SOC displayed disproportionately high intrachromosomal contacts (~20% of total; Fig. 4; Supplementary Fig. 5).

BOC1 forms a clear topologically associated domain (TAD)-like structure (Fig. 4b, c; Supplementary Fig. 6; Extended Data 3), bounded by sharp domain borders and containing four putative chromatin loops (α, β, γ, δ). Genes within these loops encode diverse functions, including glycolipid biosynthesis and chromatin binding (Extended Data 3). SOC1 exhibits a similar but smaller insulated pattern (Fig. 4d). Hi-C map of SOC suggests a TAD-like structure encompassing SOC1, though its size limits domain identification. Quantitative analysis showed that BOC and SOC have significantly fewer interchromosomal contacts than standard regions (Fig. 4e;

p<0.0001; Extended Data 3). Instead, these outlier regions preferentially interact with each other, suggesting regulatory isolation.

This spatial architecture appears to be a conserved feature of Mamiellales genome organisation. Analysis of *M. tinhauana* Hi-C data[37] revealed similar compartmentalisation patterns (Supplementary Fig. 1; Extended Data 3), suggesting that spatial insulation of outlier regions evolved early and has been maintained across the lineage.

Outlier regions maintain distinctive epigenetic landscapes

Physical compartmentalisation raises the question of how distinct gene expression is maintained in outlier regions. Whole-genome methylation profiling revealed that epigenetic and spatial boundaries are precisely aligned, providing a molecular mechanism for regulatory isolation.

In *O. tauri*, a sharp shift in methylation at ~0.55 Mb on chromosome 2 precisely demarcates the BOC1 region (Fig. 4b), aligning with the TAD-like boundary identified by Hi-C and the zone of higher TE density and gene expression. This boundary separates regions with fundamentally different chromatin states, with methylation levels within outlier regions being consistently lower than in the standard genome. BOC1 shows a 1.4-fold reduction in CpG 5-methylcytosine (5mCG) relative to the rest of the genome (Fig. 4b; Supplementary Figs. 8, 9; Extended Data 4). Overall, 5-methylcytosine (5mC) and N6-methyladenine (6mA) are reduced by 1.7- and 1.4-fold, respectively, while 5-hydroxymethylcytosine (5hmC) and 5hmCG remain uniformly low across the chromosome. SOC1 displays similarly distinctive patterns, with a sharp rise in 6mA at its start and increased 5mCG at its end (Supplementary Fig. 10).

These hypomethylated states are characteristic of permissive chromatin and provide a mechanism for the higher gene expression observed in outlier regions. Available bisulfite sequencing data[39] confirmed this pattern across species: *O. lucimarinus* shows CpG methylation declining from 22.2% in standard regions to 19.2% in BOC1 and 16.2% in SOC1. *M. pusilla* shows a similar trend with levels decreasing from 17.8% to 14.5% in BOC1 and 10.9% in SOC1 (Fig. 5a,b; Supplementary Fig. 11; Extended Data 4).

TE density modulates nucleosome spacing and chromatin flexibility

The striking differences in TE abundance between Mamiellales species provide an opportunity to test whether TE load influences chromatin structure. We therefore re-analysed published MNase-seq data from *O. lucimarinus* and *M. pusilla*[39] to determine how variation in TE density corresponds to changes in nucleosome spacing and positioning within BOC1.

Across both species, nucleosome organisation in BOC1 bears hallmarks of transcriptionally permissive chromatin: high nucleosome occupancy around transcription start sites, weaker periodicity in promoters, and reduced CpG methylation (Fig. 5c,e; Extended Data 4). These features coincide with elevated transcription:

154 RNA-Seq coverage in BOC1 is 1.86× and 2.44× higher than in standard regions for *O. lucimarinus* and *M. pusilla*,
155 respectively (Kruskal–Wallis, $p < 0.001$).

156 Nucleosome spacing differs markedly between species and genomic compartments. In the TE-rich *M. pusilla*,
157 the nucleosome repeat length shortens from 207 bp in standard regions to 196 bp in BOC1, whereas the TE-
158 poor *O. lucimarinus* shows only a minor shift from 183 bp to 181 bp (Fig. 5d,f). Shorter spacing and reduced
159 stiffness in *M. pusilla* BOC1 suggest a more dynamic chromatin fibre, potentially enhancing accessibility to
160 transcriptional machinery and TE integration sites.

161 TE–nucleosome interactions further support this model. Whereas *O. lucimarinus* BOC1 shows little TE
162 enrichment, *M. pusilla* BOC1 contains 3.5× more TEs than its standard genome and is notably depleted in
163 introner elements (IEs), which are otherwise abundant in the standard genome of this species[17,40]. This
164 contrasting distribution suggests that TEs and IEs occupy distinct chromatin niches. TEs display a bimodal
165 distribution with peaks ±50 bp from nucleosome centres, especially pronounced in BOC1, consistent with
166 preferential insertion into accessible linker DNA rather than nucleosome cores. Notably, the same regions
167 encoding giant type I PKS genes exhibit elevated MNase coverage and CpG methylation (Fig. 5a,b; Extended Data
168 4), suggesting targeted regulation of biosynthetic loci within this flexible chromatin environment.

169 BOC1 encodes diverse polyketide biosynthetic pathways with distinct metabolic 170 outputs

171 The massive TE enrichment and distinctive chromatin architecture of outlier regions are predicted to drive
172 functional diversification. Giant type I PKS genes within BOC1 provide compelling evidence for this prediction,
173 showing large structural variations and corresponding metabolic diversity across Mamiellales species.

174 PKS gene architecture varies dramatically between species, with lengths ranging from 14,393 bp in *O. tauri* to
175 64,971 bp in *M. tinhauana* (Fig. 6a; Extended Data 5). Even within species, intraspecific variation is substantial:
176 different *O. tauri* strains harbour PKS genes ranging from 14,393 to 33,947 bp. Long-read resequencing of our *O.*
177 *tauri* RCC4221 laboratory culture ("Exeter") revealed that the overall BOC1 architecture remains conserved
178 relative to the reference genome, but the PKS locus displays substantial structural reorganisation, with additional
179 modules, domain rearrangements, and multiple inversions (Fig. 6a; Extended Data 5). These features provide
180 direct evidence of ongoing genomic rearrangement and modular evolution within BOC1.

181 To explore whether these structural variations might correspond to metabolic divergence, we performed
182 untargeted metabolomics on *O. tauri* (RCC4221) and *O. lucimarinus* (CCE9901), two species with contrasting
183 PKS organisations in BOC1 (Supplementary Fig. 14; Extended Data 6). Across both ionisation modes, we
184 detected 2,213 unknown mass features, with 273 and 50 showing significantly higher abundance in *O. tauri* and

185 81 and 41 in *O. lucimarinus* in positive and negative ion modes, respectively (ANOVA, $p<0.05$, $|LFC|>2$; Fig. 6b;
186 Extended Data 6). These comparisons reveal broad chemical divergence between the species rather than
187 condition-specific regulation. Among the discriminant features, we identified one compound (RT 14.38 min, m/z
188 739.437; Supplementary Fig. 15) whose fragmentation pattern is consistent with polyketide-like structures,
189 though structural confirmation and direct demonstration of PKS involvement remain to be established.

190 Outlier regions in a phylogenetically distant picoalga

191 To explore whether specialised genomic regions are unique to Mamiellales or represent a broader evolutionary
192 strategy, we examined genomic data from another picoeukaryote with available Hi-C and RNAseq data. We
193 identified similar outlier regions in the distantly related stramenopile *Pelagomonas calceolata*[41](Supplementary
194 Fig. 16). Each chromosome in *P. calceolata* contains an outlier region (~500 kb) with ~15% lower GC content
195 compared to the rest of the genome, mirroring the pattern observed in Mamiellales. Hi-C analysis confirmed that
196 these regions are spatially compartmentalised from the standard genome[41], exhibiting the same insulation
197 architecture observed in Mamiellales BOC1 and SOC1.

198 Genes within *P. calceolata* outlier regions show ~3-fold elevated expression compared to standard regions,
199 nearly identical to the transcriptional overactivity observed in Mamiellales. Remarkably, one *P. calceolata* outlier
200 region contains a giant PKS gene positioned adjacent to a MULE element (Supplementary Fig. 17).

201 Discussion

202 Mamiellales, among the smallest free-living eukaryotes, harbour unusually large and dynamic chromosomal
203 regions, BOC1 and SOC1, with distinctive genomic and transcriptomic signatures. Genes on these regions are
204 more highly expressed than those in the standard genome across Mamiellales. This pattern mirrors land plants
205 such as *Oryza sativa* and *Zea mays*[42], where low-GC regions correlate with increased transcription, but contrasts
206 with the freshwater green alga *C. reinhardtii*[35]. Reduced 5mCG and distinctive nucleosome organisation typical
207 of open chromatin[43–45] likely underpin the transcriptional hyperactivity in these regions.

208 Hi-C analyses revealed that BOC1 and SOC1 form distinct, self-interacting domains resembling TADs. This
209 compartmentalisation likely contains TE activity while coordinating the expression of genes involved in
210 specialised pathways, including polyketide biosynthesis. These findings are consistent with earlier work showing
211 that the SOC is hypervariable in *O. tauri* and mediates antiviral defence[27,28].

212 Cryo-electron tomography in *O. tauri* previously revealed disordered higher-order chromatin lacking compact
213 fibres[46,47], interpreted as evidence for weak nucleosome phasing. By contrast, our re-analysis of MNase-seq data

214 from *O. lucimarinus* and *M. pusilla*[39] shows well-phased arrays, indicating that chromatin organisation across
215 Mamiellales is species- and locus-specific rather than uniformly diffuse.

216 The more dynamic nucleosome positioning of BOC1 and SOC1 likely facilitates both high gene expression and
217 local TE mobilisation[48,49]. The more closely spaced nucleosomes suggest a regulated nucleosome arrangement
218 that allows access to transcriptional machinery and TE insertion sites[48,49]. However, the relationship between
219 chromatin architecture and TE abundance varies markedly across Mamiellales lineages. In TE-rich *M. pusilla*,
220 the bimodal distribution of TEs around nucleosome centres supports a regulatory role of chromatin organisation
221 in TE activity and rearrangement, whereas *O. lucimarinus* retains similar chromatin features despite minimal TE
222 content, and *B. prasinos* lacks both abundant TEs and PKSs yet shows elevated expression. Notably, introner
223 elements, intron-resident mobile elements abundant in the standard genome of *M. pusilla*[17,40], are largely absent
224 from BOC1, suggesting that their insertion requires more compact chromatin, whereas canonical TEs thrive in
225 the open architecture of outlier regions.

226 In *Arabidopsis* the histone variant H2A.Z directs Ty1/Copia insertions toward stress-responsive genes while
227 protecting essential housekeeping loci from disruptive integration[50]. Because Mamiellales also encode an H2A.Z
228 homolog, a comparable mechanism may confine TEs to BOC1 and SOC1. In *O. tauri*, this includes Ty1/Copia
229 families enriched in BOC1, creating a controlled space for adaptive innovation without compromising essential
230 functions. In TE-poor species like *B. prasinos*, the same chromatin architecture may instead promote innovation
231 through alternative mechanisms such as alternative splicing, which is more frequent in BOC1 and SOC1, as also
232 observed in *M. pusilla* and *M. commoda*[51]. In these compact genomes, splicing offers a route to proteomic
233 diversification. The interplay among TE activity, chromatin dynamics, and splicing regulation likely enhances the
234 functional versatility of genes in BOC1 and SOC1.

235 PKSs exemplify the functional payoff of outlier chromosome evolution. These giant modular enzymes generate
236 a vast array of secondary metabolites (including antibiotics, toxins, and signalling molecules) and their
237 diversification is a hallmark of adaptive metabolic innovation[52,53]. In Mamiellales, the presence of PKS loci within
238 BOC1 links TE-rich genomic architecture to chemical versatility.

239 TEs are powerful engines of genomic innovation, enabling gene duplication, rearrangement, and domain
240 shuffling[54]. In the compact, haploid genomes of Mamiellales, where homologous recombination is limited[38], TE
241 activity near PKS loci provides one of the few mechanisms for expanding biosynthetic capacity. We observed
242 repeated associations between MULEs and PKS genes in *O. tauri* and other Mamiellales taxa. Because MULEs
243 promote genomic rearrangements and gene duplications[55], their proximity to PKS genes likely accelerates
244 diversification of module architecture, expanding polyketide repertoires and generating novel adaptive
245 traits[52,56,57].

246 Genome resequencing of *O. tauri* revealed structural variation within PKS loci, and untargeted metabolomics
247 detected species-specific chemical differences consistent with metabolic diversification. Among these, one
248 feature, differing in abundance between *O. tauri* and *O. lucimarinus*, showed properties compatible with a
249 polyketide-like compound, though its structure and origin remain unconfirmed. In synthetic biology, "plug-and-
250 play" engineering of modular PKSs exploits recombination boundaries defined by conserved motifs that preserve
251 catalytic efficiency across otherwise incompatible modules[58]. By analogy, MULE activity near PKS loci in
252 *Ostreococcus* may leverage these naturally permissive junctions *in vivo*, as synthetic biologists do *in vitro*,
253 turning outlier regions into natural platforms for molecular innovation.

254 Importantly, TE-PKS associations are not unique to Mamiellales. In the stramenopile picoalga *Pelagomonas*,
255 low-GC outlier regions also show elevated expression, TE insertions, and PKS expansion, and Hi-C analysis
256 confirms that these regions are spatially compartmentalised from the standard genome, exhibiting the same
257 insulation architecture observed in Mamiellales BOC1 and SOC1. This parallel organisation implies a broader
258 role for TE-mediated and chromatin-constrained metabolic innovation in marine phytoplankton. The ecological
259 significance of PKSs is evident across diverse algae: in *C. reinhardtii*, a giant type I PKS supports zygospore
260 maturation and stress tolerance[59], while in the haptophyte *Prymnesium parvum*, massive "PKZILLA" enzymes
261 drive toxin biosynthesis conferring ecological competitiveness[53]. By enabling genomic rearrangements that alter
262 PKS module composition, TEs facilitate combinatorial diversity in secondary metabolism, potentially enhancing
263 responses to environmental challenges. Thus, PKS loci illustrate how genome architecture and TE dynamics
264 converge to generate biochemical diversity. Through TE-mediated rearrangements and modular recombination,
265 outlier regions serve as experimental platforms for secondary metabolism, through which streamlined algae
266 continuously reinvent their chemistry to meet shifting ecological pressures.

267

268 Although direct evidence of sexual reproduction in Mamiellales is lacking, genomic features suggest it may occur
269 and influence chromosome architecture[29,60]. In *O. tauri*, BOC1 combines suppressed recombination with dense
270 TE accumulation, which are hallmarks of mating-type loci in other eukaryotes[61]. Yet some lineages, such as *B.*
271 *prasinos*, exhibit TE-poor BOC1 regions lacking PKS genes, resembling PKS deletions reported in some *O. tauri*
272 strains[56]. This diversity reflects lineage-specific histories of TE activity and gene turnover with adaptive
273 consequences. Consistent with this view, population analyses of *Bathycoccus* indicate the persistence of
274 outlier-linked haplotypes segregating in natural populations[19]. Collectively, our findings support a model in which
275 outlier regions act as innovation hubs that leverage transcription hyperactivity, splicing, and 3D insulation to
276 promote metabolic diversification. When TEs are abundant, they amplify and accelerate diversification,
277 especially at PKS loci, whereas non-TE processes (e.g., microhomology-mediated repair or segmental

duplication) provide slower structural change. Our analyses reveal that BOC1 and SOC1 remain transcriptionally active and hypomethylated despite containing TEs and exhibiting reduced recombination. Their insulation and unique composition suggest they function as supergenes, regions where adaptive genes and TEs co-accumulate under specialised regulatory control[62]. This organisation confines TE activity to controlled compartments, allowing mobilisation while shielding the core genome. This architecture contrasts with other eukaryotic models where TE accumulation, low GC content, and recombination suppression typically mark centromeric regions or, in fungi, repeat-induced point mutation (RIP). In these models, chromatin is repressive and hypermethylated, leading to transcriptional silencing and mutation of repetitive sequences[63,64].

Taken together, BOC1 and SOC1 emerge as conserved innovation hubs across Mamiellales. Their low-GC composition, chromatin features, and spatial insulation provide the framework for TEs to drive rearrangements, alternative splicing, and metabolic diversification. Although their full functional repertoire remains to be characterised, enrichment in secondary-metabolism and antiviral genes suggests key roles in environmental resilience. The link between SOC size and viral resistance in *Ostreococcus*[27], and parallels in *Pelagomonas*, support a general model in which outlier chromosomes act as testbeds for TE-driven innovation in minimal eukaryotic genomes, sustaining adaptability in dynamic ocean ecosystems[2,15,16,65].

Methods

Algal strains and culturing

A total of eight Mamiellales species were included in the study, representing the genera *Ostreococcus*, *Bathycoccus*, *Micromonas* and *Mantoniella*. We used *Ostreococcus tauri* strain RCC4221 in this study to generate data for Hi-C, RNA-Seq and methylation profiling. We sourced other molecular datasets from publicly available databases. The species, strain IDs and datasets used in this study are provided in Supplementary Table 1.

O. tauri strain RCC4221 was maintained in L1 seawater medium[66] in non-treated polycarbonate flasks and maintained in a 14 h: 10h light: dark cycle at 35 µmol m^{-2} s^{-1} irradiance and a temperature of 20 °C. To achieve a near-axenic culture of *O. tauri*, we isolated cells by flow cytometric cell sorting (BD Influx, Beckton Dickinson) and maintained algal cells in exponential growth by weekly subculturing. We also used a monthly antibiotic treatment (5 µg Ampicillin, 1 µg Gentamicin, 2 µg Kanamycin, 10 µg Neomycin). Algal growth was regularly monitored by flow cytometry (Cytoflex, Beckman Coulter) based on side-angle light scatter and chlorophyll red fluorescence. Bacterial cells were monitored and quantified via flow cytometry after SYBR Green I staining.

Transcriptomics

RNA extraction and sequencing. *O. tauri* RCC4221 cells were collected by centrifuging a 20 mL sample (10 min at 4000 × g at 4 °C), snap-frozen in liquid nitrogen, and stored at -80 °C. RNA was extracted using a RNeasy Plant kit (QIAGEN) with RLT buffer incubated at 56 °C for 3 min. The residual DNA was digested using the Turbo DNA-free kit (Thermo Fisher Scientific). RNA was quantified using a Qubit Fluorometer (Thermo Fisher Scientific), and RNA quality was assessed using an Agilent 2100 Bioanalyzer System. RNA extracts with an RNA integrity number (RIN) greater than eight were subjected to poly(A) selection prior to library preparation with the Illumina TruSeq Direction Library Prep Kit and sequenced on Illumina NovaSeq 6000 (150 bp paired end reads).

RNA-Seq data processing. Adapters and low-quality reads from RNA-Seq libraries were trimmed using Fastp v0.23.2[67] (minimum length required 50). Quality control reports were compiled using FastQC v0.11.9[68], MultiQC v1.13[69], and fastp. Ribosomal RNA reads were removed using SortMERNA v4.3.6[70] matching reads with rRNA sequences from the SILVA databases[71]. The remaining reads were aligned and quantified to the *O. tauri* reference genome sequence (GCF_000214015.3) using Star v2.7.10[72] with default parameters. Counts were divided by gene length to standardise counts and generate transcript levels. For publicly available RNA-Seq libraries and counts, transcript levels were measured in *M. pusilla* CCMP1545 using table 'Sup_Table_S2_CCMP1545' column 'P-Replete T1.5 hours' for counts[73]. For *M. commoda* RCC299, the existing triplicate axenic counts[74]. For *B. prasinos* RCC1105, *O. lucimarinus* CCE9901, *M. commoda* RCC299 the reads from SRR1300453[75], SRR847308[39] and SRR17220685[74] were used with the same workflow as above to additionally measure splice events with Star.

Oxford Nanopore Technology sequencing

DNA extraction and library preparation. Cells were collected by centrifuging a 30 mL sample (10 min at 4000 × g at 4 °C), snap-frozen in liquid nitrogen, and stored at -80 °C. DNA was extracted using a DNeasy Plant kit (QIAGEN). DNA was quantified using a Qubit Fluorometer (Thermo Fisher Scientific). Library preparation was conducted using the Ligation Sequencing Kits (SQK-LSK109; SQK-LSK114; ONT, Oxford, UK). Genomic DNA was pooled for sequencing on an FLO-FLG001 Flongle flow cell and FLO-MIN114 Minion flow cell (R9.4.1; R10.4.1; ONT, Oxford, UK).

Methylation profiling. Super-accuracy base calling was performed on R10.4.1 sequence data using Dorado v0.6.0 (ONT, Oxford, UK), Methylation profiling included N6-methyladenine (6mA), N5-methylcytosine (5mC/5mCG), and N5-hydroxymethylcytosine (5hmC/5hmCG) which were transformed into BED files using modkit v0.2.8 (ONT, Oxford, UK). To normalise for variations in local GC content, % methylation levels represent the average proportion of the methylated base relative to the total number of that base within the sliding window.

Genome assembly. Exeter RCC4221 was de novo assembled from R9.4.1 ONT sequence data using Flye v2.9.2[76] (--iterations 3 –scaffold) followed by medaka_consensus v1.8.0 (ONT, Oxford, UK). Matching scaffolds to reference were pulled out and aligned using blastn v2.1.0[77].

High-throughput chromosome conformation capture

Hi-C sequencing. For consistency, samples for all Hi-C experiments were collected at 2 hours into the light phase. To ensure sufficient DNA yield for sequencing from small *O. tauri* genomes, 1–3 × 10^8 total cells were used in all experiments. As there was no published protocol for picoeukaryotes we tested both *in situ*[78] and *in vivo* Hi-C approaches and a series of optimisation steps, which are detailed in the Supplementary Materials. In summary, the final protocol was as follows: we performed crosslinking by adding 2% v/v final concentration formaldehyde (16% TEM grade, Sigma) directly within the culture medium containing the cells and incubating at room temperature for 20 mins. Cross-linking was stopped by adding glycine (0.2 M final concentration) and incubating for 5 mins at room temperature on a roller. Finally, cells were washed with ice-cold PBS and stored at -80 °C. The final Hi-C library was generated using the Proximo Microbe Hi-C kit v2 (PhaseGenomics, USA) following the manufacturer's instructions and sequenced using the Illumina NovaSeq 6000 platform (150 bp paired end reads) yielding a total of 94143446 reads (run ID V0017_combined in Extended Data 3).

Hi-C data processing. Adapters and low-quality reads were trimmed from Hi-C libraries using Fastp v0.23.2. Quality control reports were compiled using FastQC v0.11.9[68], MultiQC v1.13[69], and fastp. The Hi-C processing pipeline, with the restriction sites of "^GATC,^AATT" and ligation sites of "GATCGATC,GATCAATT,AATTGATC,AATTAATT", was based on NextFlow v22.04.5 (available as a Docker profile: nf-core/hic v1.3.0)[79–81]. Hi-C reads were mapped onto a soft and hard-masked (Supplementary Fig. 18). *O. tauri* reference genome sequence (GenBank identifier: GCF_000214015.3) using Bowtie v2.3.5.1[82]. Duplication removal and contact pair validation were conducted with HiC-Pro v2.11.1[79]. Raw and normalised contact maps were generated using cooltools v0.3.0[83] at bin resolutions of 5, 10 and 20 kb. Additional normalisation was carried out using 'iced' with HiC-Pro, a Python implementation of the ICE normalisation algorithm from Cooltools[84]. Hi-C quality reports were generated using HiCUP[85]. Domain compartment calling was completed using Cooltools. The R package HiCRep[86] was used to evaluate the correlation between Hi-C experimental replicates (Supplementary Fig. 19). Juicer v1.6[87] was used to produce additional contact maps for use with downstream loop detection. Chromatin loops were identified from the Nextflow and Juicer contact maps using Mustache v1.0.1[88] and HiCCUPS v3.0[87,88], respectively, at bin resolutions of 5, 10 and 20kb., its derivative/slope and cis and trans interactions was calculated using cooltools. *M. tinhauana* RCC11003 Hi-C data[37] was similarly analysed[37].

Additional sequence analyses

MNase-seq analysis. MNase-seq data for *M. pusilla* CCMP1545 and *O. lucimarinus* CCE9901 (SRR847304 and SRR847307[39]) were processed to analyse nucleosome positioning. Paired end reads underwent quality trimming and adapter removal using fastp v0.23.4. The cleaned reads were aligned to the reference genome with Minimap2 v2.24[89] using default parameters for paired-end data. Aligned files were converted to BAM format, sorted, and indexed with Samtools v1.13[90]. Finally, genome-wide coverage was calculated and normalised to RPKM using bamCoverage (deeptools v3.5.1[91], --MNase --center-reads) to calculate nucleosome centers, and the results were exported as bigwig files for visualisation and downstream analysis in Python for peak calling. Autocorrelation functions of nucleosome center coverage were estimated using python. Nucleosome fuzziness and stiffness were calculated using Nucleosome Dynamics v1.0[92].

Bisulfite-seq analysis. Bisulfite sequencing data for *M. pusilla* CCMP1545 and *O. lucimarinus* CCE9901 (SRR847303 and SRR847306[39]) were processed to determine methylation levels at CpG sites. Single-end reads were trimmed for quality and adapters using Fastp v0.23.4. The trimmed reads were aligned to the reference genome using Bismark v0.24.2[93]. Methylation calling was restricted to CpG contexts, and methylation levels were summarised in bedgraph and bigwig format for further analysis.

Homology and GO analyses. Protein synteny was highlighted using Oxford dot plots v0.3.3[94], using blastp v2.10[77] and Diamond v2.1.8[95] across outlier chromosomes in the Mamiellophyceae. A ribbon plot was generated to show the synteny between species using Python. EDTA v.2.2.2[96], default settings, were used to flag transposable elements across whole genomes and PKS genes, respectively. Enrichment of TEs was calculated using 5 kb bins and Kolmogorov–Smirnov test. A data frame of all the proteins within *O. tauri* with GO enrichment spanning biological processes, molecular functions, cellular components, and locations on chromosomes was obtained from NCBI and UniProt[97]. Fisher's exact test (one-tailed) was performed using a custom script in Python on subsets of genes to compare enriched terms over target regions and the rest of the genome.

Untargeted metabolomics

Metabolite profiling was carried out using LC-quadrupole time of flight mass spectrometry using separate runs for positive and negative ion modes. *O. tauri* RCC4221 and *O. lucimarinus* CCE9901 cells were collected by centrifuging a 30 mL sample (10 min at 4000 × g at 4 °C), snap-frozen in liquid nitrogen, and stored at -80 °C. Pellets were extracted in 0.6 ml methanol, treated for 20 min in a sonicating water bath with ice and then centrifuged at 20,000g for 10 min at 4 °C. The supernatants from three replicate cultures of each species and extraction blanks were analysed using an Agilent Zorbax Eclipse plus, narrow bore RR, 2.1 × 50 mm, 3.5 µm

402 particle size reverse phase C18 analytical column (Agilent Technologies, Palo Alto, USA). 5 µL samples were
403 injected *via* an Agilent 1200 series Rapid Resolution HPLC system coupled to a quadrupole time-of-flight (QToF)
404 6520 mass spectrometer (Agilent Technologies, Palo Alto, USA). Mobile phases were A (10% methanol, 0.1%
405 formic acid), mobile phase B (100% methanol with 0.1% formic acid). The following gradient was used: 0–5 min-
406 30% B; 10–24 min-0–98% B; 25-30 min-100% B. The flow rate was 0.3 mL min^{-1} and the column temperature
407 was 40 °C. The source conditions for electrospray ionisation were as follows: gas temperature was 325 °C with
408 a drying gas flow rate of 9 l min^{-1} and a nebuliser pressure of 35 psig. The capillary voltage was 3.5 kV in both
409 positive and negative ion mode. The fragmentor voltage was 175 V and skimmer 65 V. Scanning was performed
410 using the auto MS/MS function at 5 scans s^{-1} for precursor ion surveying and 4 scans s^{-1} for MS/MS with a sloped
411 collision energy of 3.5 V/100 Da with an offset of 10 V. MS and MS/MS m/z range was 40-1700. Reference
412 compound m/z values were 121.05087 and 922.00979. Raw data files (*.d) were converted to mzML format
413 using MSConvert (ProteoWizard)[98]. Feature detection and alignment was carried out with MS-Dial v5[99]. Features
414 present in the blank (>20% of sample intensity) were discounted. Compounds were identified with the MS-Dial
415 lipidomics workflow and using publicly available CompMS MS-Dial MS/MS libraries. Statistical one-factor Anova
416 test analysis ($p<0.05$, $|LFC|>2$) was then performed on 'Unknown' features with Metaboanalyst v6.0[100] to being
417 used highlight features for further annotation for putative PKS products within Sirius[101].

418 ## Supplementary Information

419 In supplementary files.

420 ## Data availability

421 Sequencing data have been deposited at NCBI under BioProject accession PRJNA1250464. Additional data
422 including Hi-C contact matrices and metabolomic profiles are available at Zenodo: 10.5281/zenodo.14284467.

423

424 ## Acknowledgments

425 Illumina sequencing was conducted on equipment funded by the Wellcome Trust (Multi-User Equipment Grant
426 award number 218247/Z/19/Z). We are grateful to Dr Karen Moore (Exeter Sequencing Facility, University of
427 Exeter) for their technical advice on Hi-C sequencing. We thank Dr Aurélie Chambouvet and Dr Evelyne Derelle
428 (CNRS, France) for providing algal strains. This research was supported by the Royal Society through a University
429 Research Fellowship (URF\R1\180537) and Research Fellows Enhancement Awards (RGF\EA\181012)
430 awarded to AM.

431

Figure legends

Figure 1. Comparative genomic analysis of BOC1 regions across Mamiellales species.

(**a**), Syntenic analysis and gene homology visualisation of BOC1 regions across six Mamiellales species. Connecting lines indicate homologous genes identified through reciprocal best BLASTP hits between *O. tauri* RCC4221, *O.* sp. RCC809, *O. lucimarinus* CCE9901, *B. prasinos* RCC1105, *M. pusilla* CCMP1545, and *M. commoda* RCC299. BOC1 regions are displayed in green. BOC1-localised orthologues are indicated by green lines. Total BOC1 length and the proportion contained within the region (in parentheses) are indicated for each species. (**b**), Gene Ontology (GO) enrichment analysis of BOC1-encoded proteins in RCC4221. The plot displays GO terms categorised into three major classes: biological process (pink), cellular component (blue), and molecular function (yellow). Enrichment values are shown as $-\text{Log}_{10}$(p-value), with a significance threshold of $p<0.0335$ (Fisher's exact test). Dot size indicates gene count, while colour intensity represents the odds ratio, defined as the proportion of GO term occurrence in BOC1 proteins relative to non-BOC1 regions.

Figure 2. Distribution and expression of transposable elements across Mamiellales genomes.

(**a**), Dendrogram of major transposable element (TE) classes present in Mamiellales. The left sidebars indicate the classification of elements into retrotransposon, with further subdivision into LTR/non-LTR retrotransposons and DDE/D transposons, rolling-circle elements, and YR cryptons. Heatmap shows TE density (TEs MB^{-1}) across standard (St), BOC1 (B), and SOC1 (S) genomic regions. The lower dendrogram displays phylogenetic relationships between Mamiellales species: *O. tauri* (Ot; strains RCC4221 and RCC1115), *O.* sp. RCC809 (Os), *O. lucimarinus* (Ol; CCE9901), *B. prasinos* (Bp; RCC1105), *M. pusilla* (Mp; CCMP1545), *M. commoda* (Mc; RCC299), and *M. tinhauana* (Mt; RCC11003). (**b**), Quantitative comparison of TE abundance (TEs MB^{-1}) across different genomic regions. Bar colours correspond to standard (grey), BOC1 (green), and SOC1 (red) regions. (**c—f**), RNA-Seq expression profiles of representative TEs in *O. tauri* RCC4221 showing: (c) Ty1/Copia element localised upstream of a putative hexosyltransferase gene; (d) Tc1/Mariner element near a transcription factor IIA; (e and f) Mutator elements near polyketide synthases. Grey peaks represent RNA-Seq coverage, with maximum coverage values indicated in the top right corner of each track. Dotted lines represent 200x RNA-Seq coverage threshold. Grey bars indicate predicted protein-coding regions with their corresponding NCBI locus tags (genome assembly GCF_000214015.3). Updated protein annotations were determined using AlphaFold3[102] and FoldSeek[103] analysis.

Figure 3. Higher gene expression in BOC1.

(**a—c**), Expression levels, GC content and TE content across BOC of (**a**) *O. tauri* (RCC4221; chromosome 2), (**b**), *M. commoda* (RCC299; chromosome 1), and (**c**) *B. prasinos* (RCC1105; chromosome 14). Boundary points of the BOC1 regions are labelled with cyan dotted lines; transposable elements are shown on the lower

track. (**d—f**), Comparative analysis of gene expression in outlier regions and standard genome; box plots display the distribution of transcript levels across outlier (BOC1 and SOC1) and standard regions for each taxon (Kruskal-Wallis, $p<0.001$). (**g—i**), Correlation between gene expression and GC content of coding sequences (CDS) in (**g**), *O. tauri* (Spearman's rank, $\rho=-0.366$, $p=7.82e^{-242}$), (**h**), *M. commoda* ($\rho = -0.285$, $p = 3.70e^{-192}$), and (**i**) *B. prasinos* ($\rho = -0.38302$, $p = 8.49e^{-13}$); CDS in BOC1 and non-BOC1 regions are represented in green and blue, respectively.

Figure 4. **Chromosomal organisation and three-dimensional genome architecture in *O. tauri*.**
(**a**), Hi-C contact map of the *O. tauri* genome binned at 10 kb resolution with normalised interactions counts; chromosomes 2 and 19, which display enriched intrachromosomal contacts, highlighted by white squares. (**b**), Hi-C normalised contact map of chromosome 2 at 20 kb resolution. The first track displays transcript levels alongside GC content. The second track shows the rolling average of DNA modifications within a scaled window, 5mCG (orange), 5mC (green), 6mA (purple), 5hmCG (blue). The third track indicates positions of transposable elements (black).
(**c—d**), Hi-C contact maps binned at 10 kb for (**c**) chromosome 2 and (**d**) chromosome 19. Topologically associating domains are indicated by black dotted lines, with outlier regions labelled as BOC1 and SOC1 (cyan text), the bottom track represents the insulation score calculated as a rolling z-score from the summation of Hi-C contacts at 5 kb (red), 10 kb (blue), and 20 kb (green) resolutions. (**e**), Distribution of normalised interchromosomal contact probabilities between different chromosomes. Interactions involving BOC or SOC chromosomes are labelled in green, whilst interactions between standard nuclear are shown in blue. Notable interactions between chromosomes 2 and 19, and between chromosomes 18 and 20 are highlighted.

Figure 5. **MNase and CpG profiles across BOC1 in *Ostreococcus lucimarinus* CCE9901 and *Micromonas pusilla* CCMP1545.**
(**a, b**), Transcript levels, CpG methylation as a percentage of base coverage which is methylated, MNase coverage alongside TEs and IEs in *O. lucimarinus* CCE9901 (**a**) and *M. pusilla* CCMP1545 (**b**). Dotted lines indicate the edge of outlier regions. Blue bars indicate PKS coding regions. (**c, e**), Means at each position aligned to transcription start sites are shown for CpG methylation and nucleosomes within the standard genome above and BOC1 genes below for *O. lucimarinus* (**c**) and *M. pusilla* (**e**). Blue is the percentage of bases which are CpG methylated. Black is nucleosome center counts. Light grey is MNase coverage. (**d, f**), The autocorrelation function estimates of nucleosome center counts are shown for each lag (offset, bp) across standard genome (black) and BOC1 (turquoise) genes for *O. lucimarinus* (**d**) and *M. pusilla* (**f**), respectively. Period is annotated in base pairs.

Figure 6. Modular Giant Type I Polyketide synthases show structural diversity across Mamiellales species.

(**a**), Schematic representation of the largest orthologous polyketide synthases identified across Mamiellales species, illustrating large size variation in protein product, from 4,794 amino acid residues in *O. tauri* RCC7083 to 20,889 residues in *M. tinhauana* driven by differences in modular domain repetition. Coloured boxes represent different functional domains within the modular PKS architecture. (**b**), Volcano plot showing differential metabolite abundances between *O. lucimarinus* (CCE9901) and *O. tauri* (RCC4221) across 1318 unidentified positive ion features, revealing 273 features enriched in *O. tauri* versus 81 enriched in *O. lucimarinus*. Fold-change thresholds and statistical significance ($p < 0.05$) are indicated by dashed lines. A putative polyketide showed 5.29-fold higher abundance in *O. tauri* compared to *O. lucimarinus* ($p = 0.0016$, ANOVA; RT 14.38 min, m/z 739.437, ID#3563). Gray dots represent non-significantly different metabolites (see Extended Data 6 and Supplementary Figs. 14-15 for additional structural and metabolomic analyses). (**c**), Structure of the putative polyketide-like compound inferred from LC–MS/MS fragmentation patterns.

References

1. Courties, C. *et al.* Smallest eukaryotic organism. *Nature* **370**, 255 (1994).
2. Chretiennot-Dinet, M. J. *et al.* A new marine picoeucaryote: Ostreococcus tauri gen. et sp. nov. (Chlorophyta, Prasinophyceae). *Phycologia* **34**, 285–292 (1995).
3. Noordally, Z. B. *et al.* A phospho-dawn of protein modification anticipates light onset in the picoeukaryote Ostreococcus tauri. *J Exp Bot* **74**, 5514–5531 (2023).
4. Corellou, F. *et al.* Clocks in the green lineage: Comparative functional analysis of the circadian architecture of the picoeukaryote ostreococcus. *Plant Cell* **21**, 3436–3449 (2009).
5. Martin, S. F. *et al.* Surviving starvation: Proteomic and lipidomic profiling of nutrient deprivation in the smallest known free-living eukaryote. *Metabolites* **10**, 273 (2020).
6. Marin, B. & Melkonian, M. Molecular Phylogeny and Classification of the Mamiellophyceae class. nov. (Chlorophyta) based on Sequence Comparisons of the Nuclear- and Plastid-encoded rRNA Operons. *Protist* **161**, 304–336 (2010).
7. Tragin, M. & Vaulot, D. Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding. *Sci Rep* **9**, 5190 (2019).
8. Yung, C. C. M., Redondo, E. R., Sanchez, F., Yau, S. & Piganeau, G. Diversity and Evolution of Mamiellophyceae: Early-Diverging Phytoplanktonic Green Algae Containing Many Cosmopolitan Species. *J Mar Sci Eng* **10**, 240 (2022).

9. Demory, D. *et al.* Picoeukaryotes of the Micromonas genus: sentinels of a warming ocean. *ISME Journal* **13**, 132–146 (2019).
10. Demir-Hilton, E. *et al.* Global distribution patterns of distinct clades of the photosynthetic picoeukaryote Ostreococcus. *ISME Journal* **5**, 1095–1107 (2011).
11. Monier, A., Worden, A. Z. & Richards, T. A. Phylogenetic diversity and biogeography of the Mamiellophyceae lineage of eukaryotic phytoplankton across the oceans. *Environ Microbiol Rep* **8**, 461–469 (2016).
12. Sands, E. *et al.* Genetic and physiological responses to light quality in a deep ocean ecotype of Ostreococcus, an ecologically important photosynthetic picoeukaryote. *J Exp Bot* **74**, 6773–6789 (2023).
13. Blanc-Mathieu, R. *et al.* An improved genome of the model marine alga Ostreococcus tauri unfolds by assessing Illumina de novo assemblies. *BMC Genomics* **15**, 1103 (2014).
14. Merchant, S. S. *et al.* The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science (1979)* **318**, 245–250 (2007).
15. Derelle, E. *et al.* Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. *Proc Natl Acad Sci U S A* **103**, 11647–11652 (2006).
16. Palenik, B. *et al.* The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci U S A* **104**, 7705–7710 (2007).
17. Worden, A. Z. *et al.* Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes micromonas. *Science (1979)* **324**, 268–272 (2009).
18. van Baren, M. J. *et al.* Evidence-based green algal genomics reveals marine diversity and ancestral characteristics of land plants. *BMC Genomics* **17**, 267 (2016).
19. Devic, M. *et al.* An INDEL genomic approach to explore population diversity of phytoplankton. *BMC Genomics* **25**, 1045 (2024).
20. Da Silva, O. *et al.* Genomic differentiation of three pico-phytoplankton species in the Mediterranean Sea. *Environ Microbiol* **24**, 6086–6099 (2022).
21. Keeling, P. J. Ostreococcus tauri: seeing through the genes to the genome. *Trends in Genetics* **23**, 151–154 (2007).
22. Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory for microbial ecology. *ISME Journal* **8**, 1553–1565 (2014).
23. Grimsley, N., Yau, S., Piganeau, G. & Moreau, H. Typical features of genomes in the mamiellophyceae. in *Marine Protists: Diversity and Dynamics* 107–127 (Springer, 2015). doi:10.1007/978-4-431-55130-0_6.
24. Moreau, H. *et al.* Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome Biol* **13**, R74 (2012).

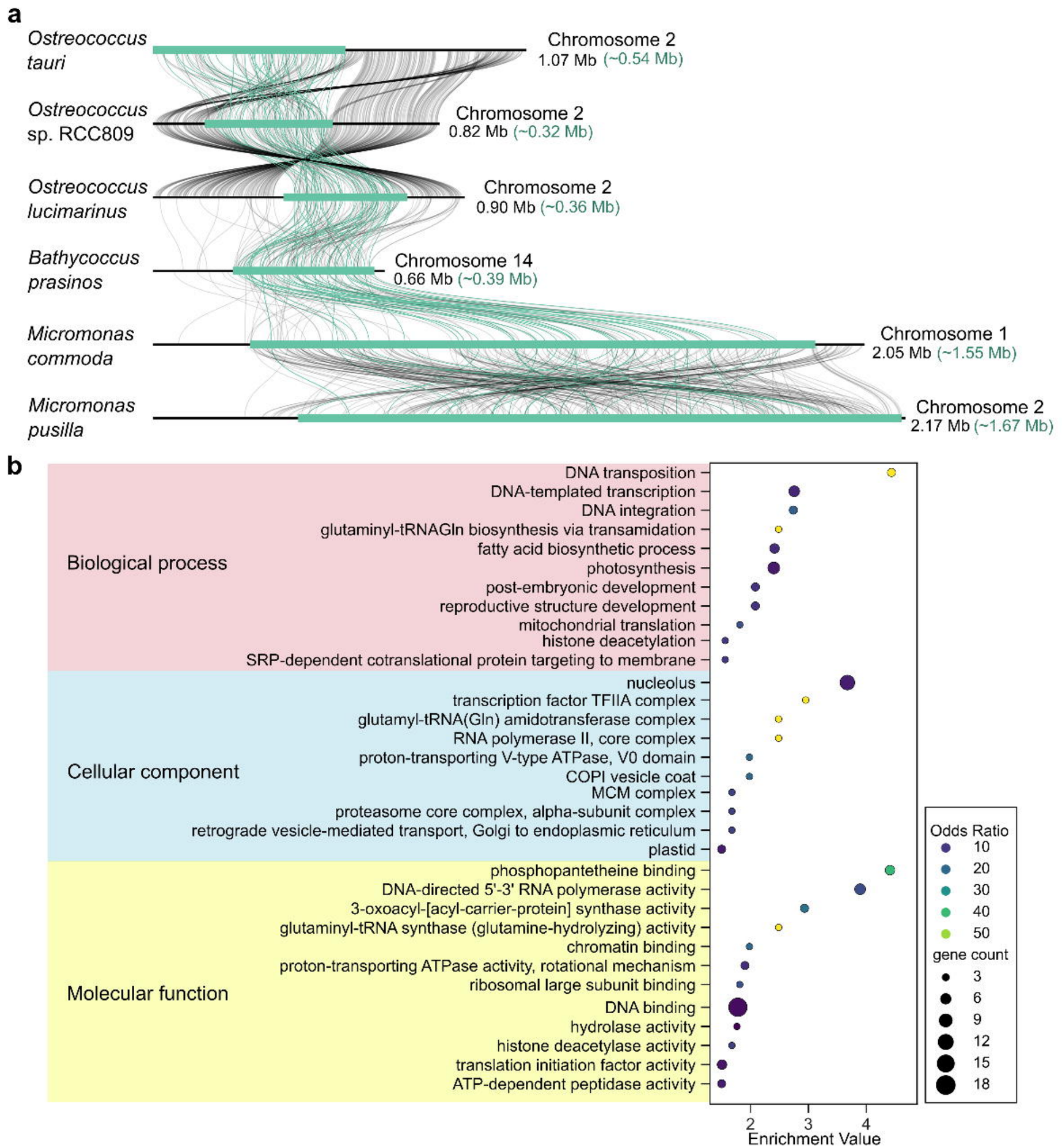
25. Lanier, W., Moustafa, A., Bhattacharya, D. & Comeron, J. M. EST analysis of Ostreococcus lucimarinus, the most compact eukaryotic genome, shows an excess of introns in highly expressed genes. *PLoS One* **3**, e2171 (2008).
26. Irimia, M. & Roy, S. W. Evolutionary convergence on highly-conserved 3' intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. *PLoS Genet* **4**, e1000148 (2008).
27. Blanc-Mathieu, R. *et al.* Population genomics of picophytoplankton unveils novel chromosome hypervariability. *Sci Adv* **3**, e1700239 (2017).
28. Yau, S. *et al.* A Viral Immunity Chromosome in the Marine Picoeukaryote, Ostreococcus tauri. *PLoS Pathog* **12**, e1005965 (2016).
29. Benites, L. F. *et al.* Evolutionary Genomics of Sex-Related Chromosomes at the Base of the Green Lineage. *Genome Biol Evol* **13**, evab216 (2021).
30. Yau, S. *et al.* Virus-host coexistence in phytoplankton through the genomic lens. *Sci Adv* **6**, eaay2587 (2020).
31. Lawson, H. A., Liang, Y. & Wang, T. Transposable elements in mammalian chromatin organization. *Nat Rev Genet* **24**, 712–723 (2023).
32. Zhang, A. & Zhang, W. Characterization of Transposon-Derived Accessible Chromatin Regions in Rice (Oryza Sativa). *Int J Mol Sci* **23**, 8947 (2022).
33. Dekker, J. GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. *Genome Biol* **8**, R116 (2007).
34. Vinogradov, A. E. DNA helix: The importance of being GC-rich. *Nucleic Acids Res* **31**, 1838–1844 (2003).
35. Barahimipour, R. *et al.* Dissecting the contributions of GC content and codon usage to gene expression in the model alga Chlamydomonas reinhardtii. *Plant Journal* **84**, 704–717 (2015).
36. Niu, D. K. & Yang, Y. F. Why eukaryotic cells use introns to enhance gene expression: Splicing reduces transcription-associated mutagenesis by inhibiting topoisomerase I cutting activity. *Biol Direct* **6**, 24 (2011).
37. Rey Redondo, E., Xu, Y. & Yung, C. C. M. Genomic characterisation and ecological distribution of Mantoniella tinhauana: a novel Mamiellophycean green alga from the Western Pacific. *Front Microbiol* **15**, 1358574 (2024).
38. Jancek, S., Gourbière, S., Moreau, H. & Piganeau, G. Clues about the genetic basis of adaptation emerge from comparing the proteomes of two Ostreococcus ecotypes (Chlorophyta, Prasinophyceae). *Mol Biol Evol* **25**, 2293–2300 (2008).
39. Huff, J. T. & Zilberman, D. Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. *Cell* **156**, 1286–1297 (2014).

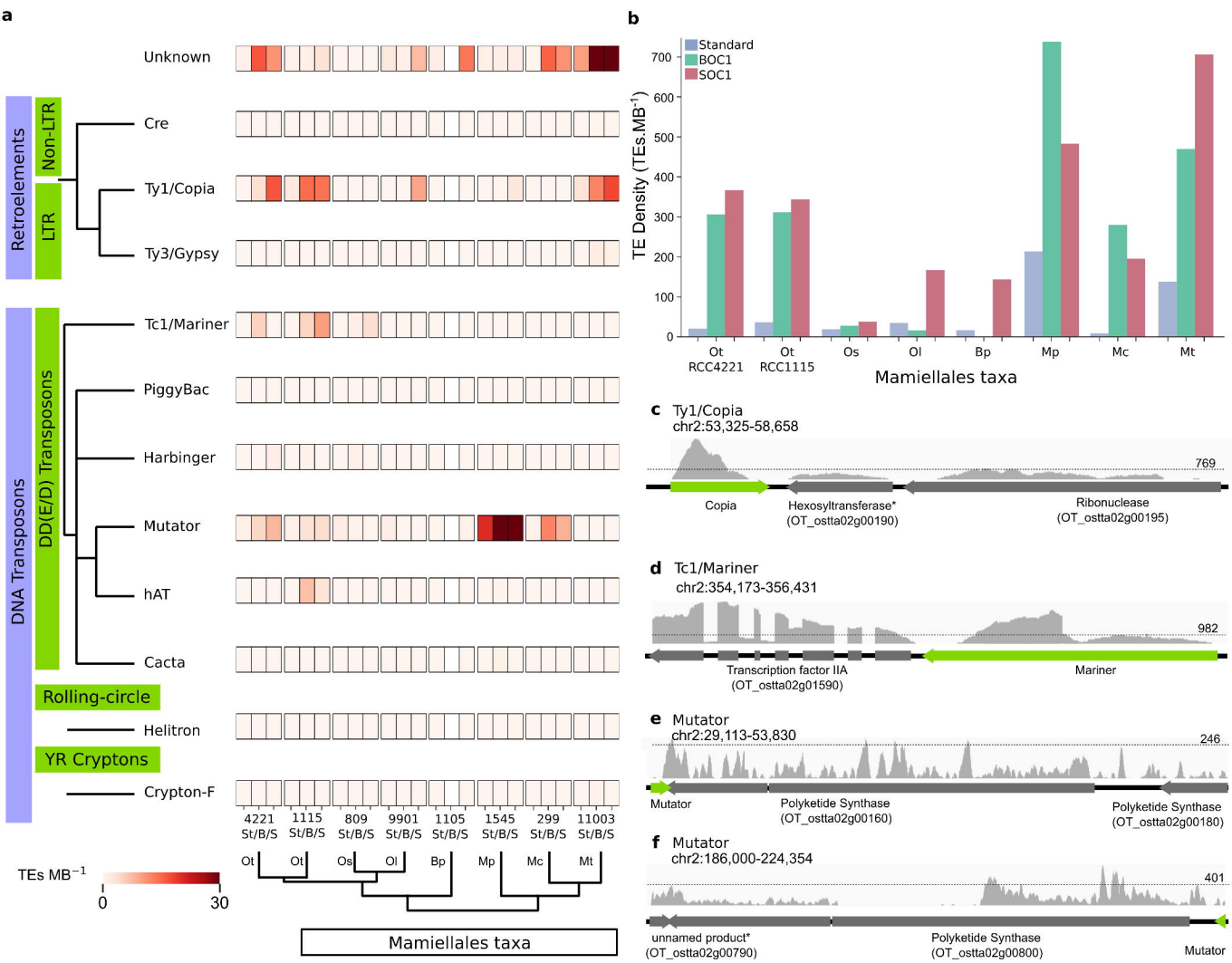
592 40. Simmons, M. P. *et al.* Intron invasions trace algal speciation and reveal nearly identical arctic and
593 antarctic micromonas populations. *Mol Biol Evol* **32**, 2219–2235 (2015).
594 41. Guérin, N. *et al.* Genomic adaptation of the picoeukaryote Pelagomonas calceolata to iron-poor oceans
595 revealed by a chromosome-scale genome sequence. *Commun Biol* **5**, 983 (2022).
596 42. Das, S. & Bansal, M. Variation of gene expression in plants is influenced by gene architecture and
597 structural properties of promoters. *PLoS One* **14**, e0212678 (2019).
598 43. Blattler, A. & Farnham, P. J. Cross-talk between site-specific transcription factors and DNA methylation
599 states. *Journal of Biological Chemistry* **288**, 34287–34294 (2013).
600 44. Kribelbauer, J. F., Lu, X. J., Rohs, R., Mann, R. S. & Bussemaker, H. J. Toward a Mechanistic Understanding
601 of DNA Methylation Readout by Transcription Factors. *J Mol Biol* **432**, 1801–1815 (2020).
602 45. Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: In the right place at the right time. *Science*
603 *(1979)* **361**, 1336–1340 (2018).
604 46. Gan, L., Ladinsky, M. S. & Jensen, G. J. Chromatin in a marine picoeukaryote is a disordered assemblage
605 of nucleosomes. *Chromosoma* **122**, 377–386 (2013).
606 47. Gan, L., Ladinsky, M. S. & Jensen, G. J. Organization of the smallest eukaryotic spindle. *Current Biology*
607 **21**, (2011).
608 48. Ohtani, H. & Iwasaki, Y. W. Rewiring of chromatin state and gene expression by transposable elements.
609 *Dev Growth Differ* **63**, 262–273 (2021).
610 49. Huda, A., Mariño-Ramírez, L., Landsman, D. & Jordan, I. K. Repetitive DNA elements, nucleosome binding
611 and human gene expression. *Gene* **436**, 12–22 (2009).
612 50. Quadrana, L. *et al.* Transposition favors the generation of large effect mutations that may facilitate rapid
613 adaption. *Nat Commun* **10**, 3421 (2019).
614 51. Verhelst, B., Van De Peer, Y. & Rouzé, P. The complex intron landscape and massive intron invasion in a
615 picoeukaryote provides insights into intron evolution. *Genome Biol Evol* **5**, 2393–2401 (2013).
616 52. Nivina, A., Yuet, K. P., Hsu, J. & Khosla, C. Evolution and Diversity of Assembly-Line Polyketide Synthases.
617 *Chem Rev* **119**, 12524–12547 (2019).
618 53. Fallon, T. R. *et al.* Giant polyketide synthase enzymes in the biosynthesis of giant marine polyether toxins.
619 *Science* **385**, 671–678 (2024).
620 54. Arkhipova, I. R. Neutral theory, transposable elements, and eukaryotic genome evolution. *Mol Biol Evol*
621 **35**, 1332–1337 (2018).
622 55. Dupeyron, M., Singh, K. S., Bass, C. & Hayward, A. Evolution of Mutator transposable elements across
623 eukaryotic diversity. *Mob DNA* **10**, 12 (2019).
624 56. Shelest, E., Heimerl, N., Fichtner, M. & Sasso, S. Multimodular type I polyketide synthases in algae evolve
625 by module duplications and displacement of AT domains in trans. *BMC Genomics* **16**, 1–15 (2015).

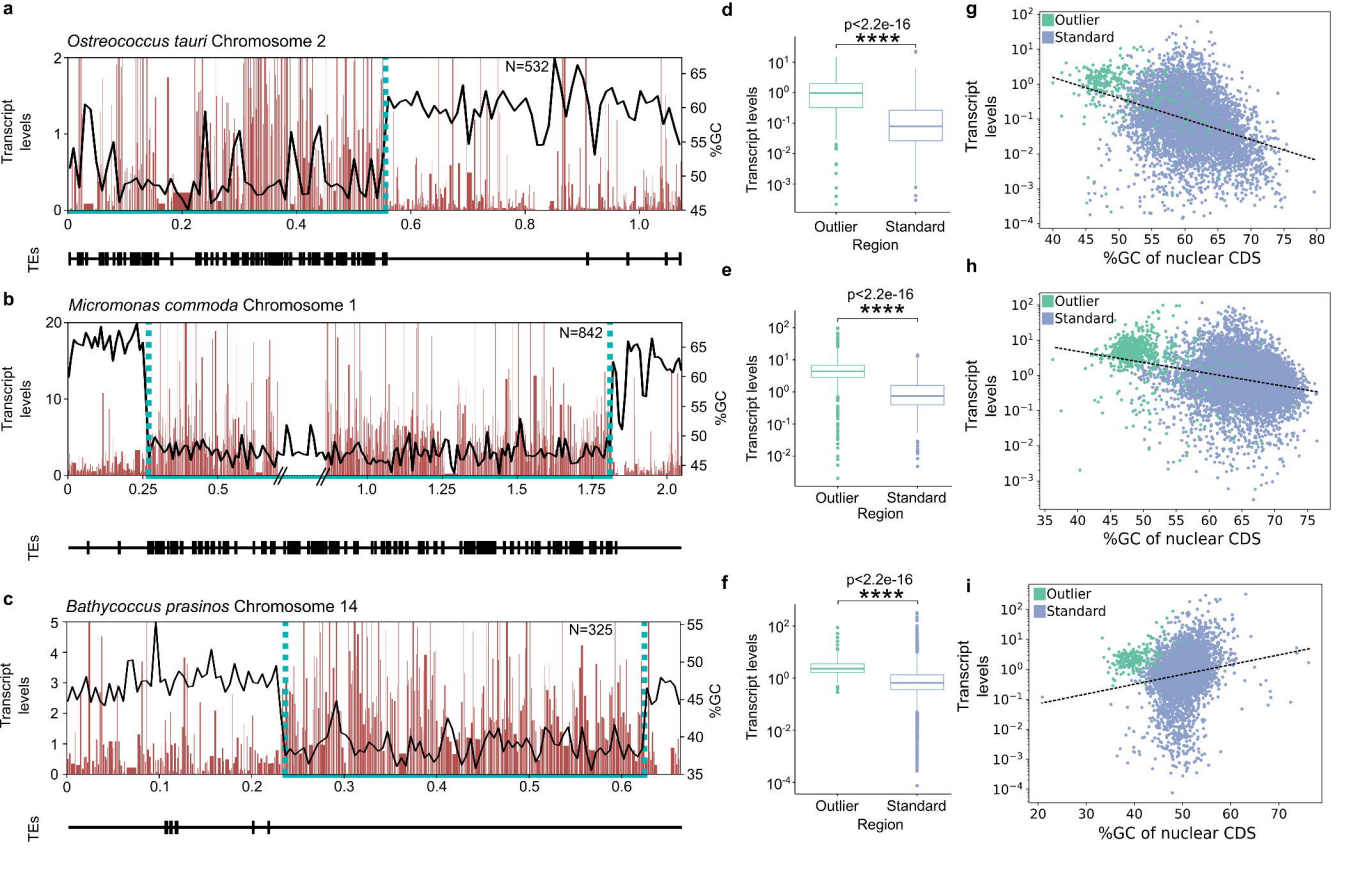
626 57. Zeng, G. *et al.* Duplication of a Pks gene cluster and subsequent functional diversification facilitate
627 environmental adaptation in Metarhizium species. *PLoS Genet* **14**, e1007472 (2018).
628 58. Huang, Z. *et al.* Plug-and-play engineering of modular polyketide synthases. *Nat Chem Biol* **21**, 1361–
629 1367 (2025).
630 59. Heimerl, N. *et al.* A giant type I polyketide synthase participates in zygospore maturation in
631 Chlamydomonas reinhardtii. *Plant J* **95**, 268–281 (2018).
632 60. Grimsley, N., Péquin, B., Bachy, C., Moreau, H. & Piganeau, G. Cryptic sex in the smallest eukaryotic
633 marine green alga. *Mol Biol Evol* **27**, 47-54. (2010).
634 61. Charlesworth, D. Evolution of recombination rates between sex chromosomes. *Philosophical*
635 *Transactions of the Royal Society B: Biological Sciences* **372**, 20160456 (2017).
636 62. Berdan, E. L. *et al.* Genomic architecture of supergenes: Connecting form and function. *Philosophical*
637 *Transactions of the Royal Society B: Biological Sciences* **377**, 20210192 (2022).
638 63. Bloom, K. S. Centromeric heterochromatin: The primordial segregation machine. *Annu Rev Genet* **48**,
639 457–484 (2014).
640 64. Irelan, J. T. & Selker, E. U. Cytosine methylation associated with repeat-induced point mutation causes
641 epigenetic gene silencing in Neurospora crassa. *Genetics* **146**, 509–523 (1997).
642 65. Henderson, G. P., Gan, L. & Jensen, G. J. 3-D Ultrastructure of O. tauri: Electron Cryotomography of an
643 Entire Eukaryotic Cell. *PLoS One* **2**, e749 (2007).
644 66. Guillard, R. R. L. & Hargraves, P. E. Stichochrysis immobilis is a diatom, not a chrysophyte. *Phycologia* **32**,
645 234–236 (1993).
646 67. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*
647 **34**, 884–890 (2018).
648 68. Andrews, S. FastQC. *Babraham Bioinformatics* (2010).
649 69. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize analysis results for multiple tools
650 and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
651 70. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in
652 metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
653 71. Glöckner, F. O. The SILVA Database Project: An ELIXIR core data resource for high-quality ribosomal RNA
654 sequences. *Biodiversity Information Science and Standards* **3**, e36125 (2019).
655 72. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
656 73. Bachy, C. *et al.* Transcriptional responses of the marine green alga Micromonas pusilla and an infecting
657 prasinovirus under different phosphate conditions. *Environ Microbiol* **20**, 2898–2912 (2018).
658 74. Ferrer-González, F. X. *et al.* Bacterial transcriptional response to labile exometabolites from
659 photosynthetic picoeukaryote Micromonas commoda. *ISME Communications* **3**, 5 (2023).

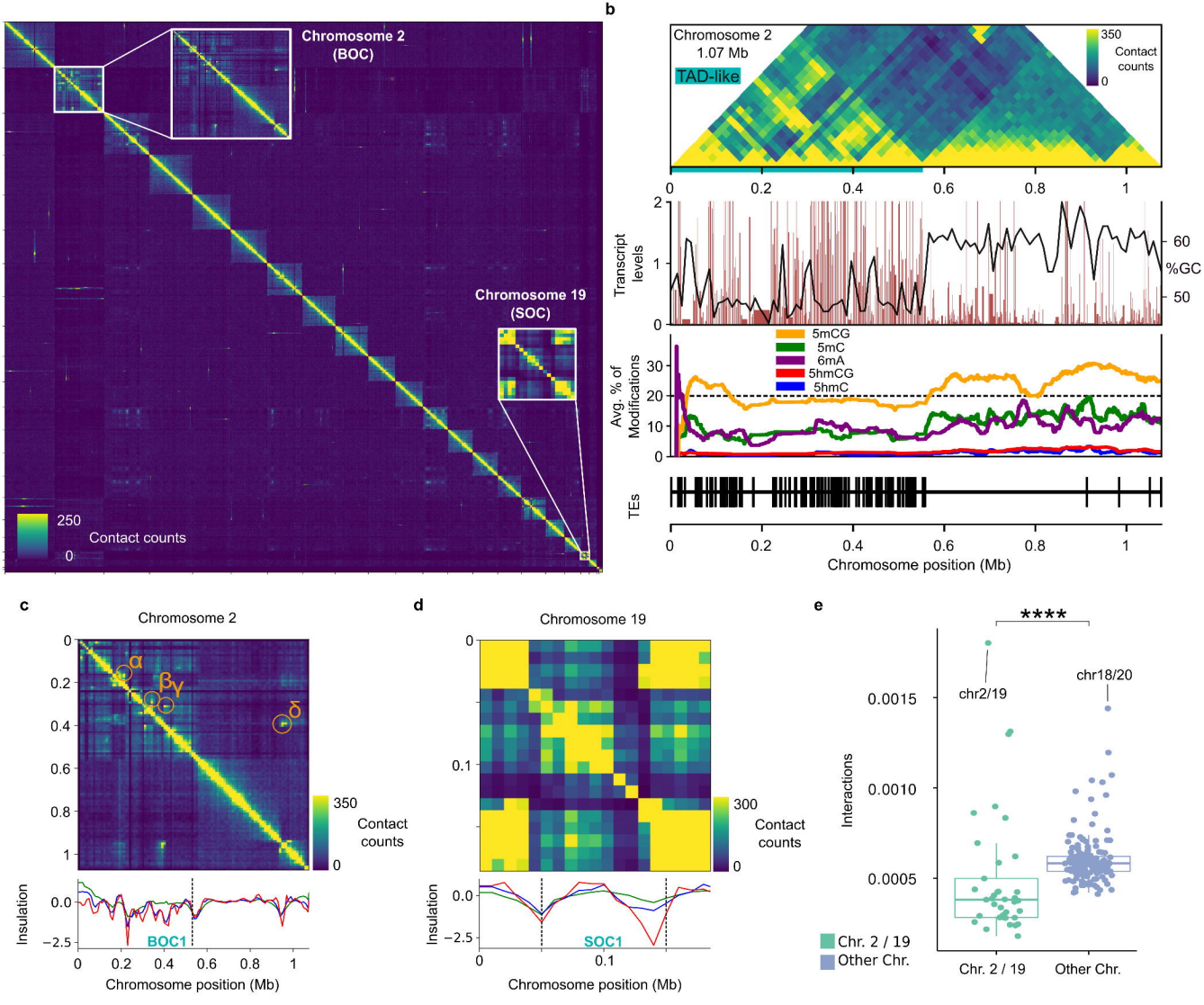
660 75. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP):
661 Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing.
662 *PLoS Biol* **12**, e1001889 (2014).
663 76. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs.
664 *Nat Biotechnol* **37**, 540–546 (2019).
665 77. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
666 78. Belton, J. M. *et al.* Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods*
667 **58**, 268–276 (2012).
668 79. Servant, N. *et al.* HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**,
669 259 (2015).
670 80. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35**, 316–
671 319 (2017).
672 81. Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol*
673 **38**, 276–278 (2020).
674 82. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359
675 (2012).
676 83. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays.
677 *Bioinformatics* **36**, 311–316 (2020).
678 84. Abdennur, N. *et al.* Cooltools: Enabling high-resolution Hi-C analysis in Python. *PLoS Comput Biol* **20**,
679 e1012067 (2024).
680 85. Wingett, S. *et al.* HiCUP: Pipeline for mapping and processing Hi-C data. *F1000Res* **4**, 1310 (2015).
681 86. Yang, T. *et al.* HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation
682 coefficient. *Genome Res* **27**, 1939–1949 (2017).
683 87. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments.
684 *Cell Syst* **3**, 95–98 (2016).
685 88. Roayaei Ardakany, A., Gezer, H. T., Lonardi, S. & Ay, F. Mustache: Multi-scale detection of chromatin loops
686 from Hi-C and Micro-C maps using scale-space representation. *Genome Biol* **21**, 256 (2020).
687 89. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
688 90. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
689 91. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. DeepTools: A flexible platform for exploring
690 deep-sequencing data. *Nucleic Acids Res* **42**, 187–191 (2014).
691 92. Buitrago, D. *et al.* Nucleosome Dynamics: a new tool for the dynamic analysis of nucleosome positioning.
692 *Nucleic Acids Res* **47**, 9511–9523 (2019).

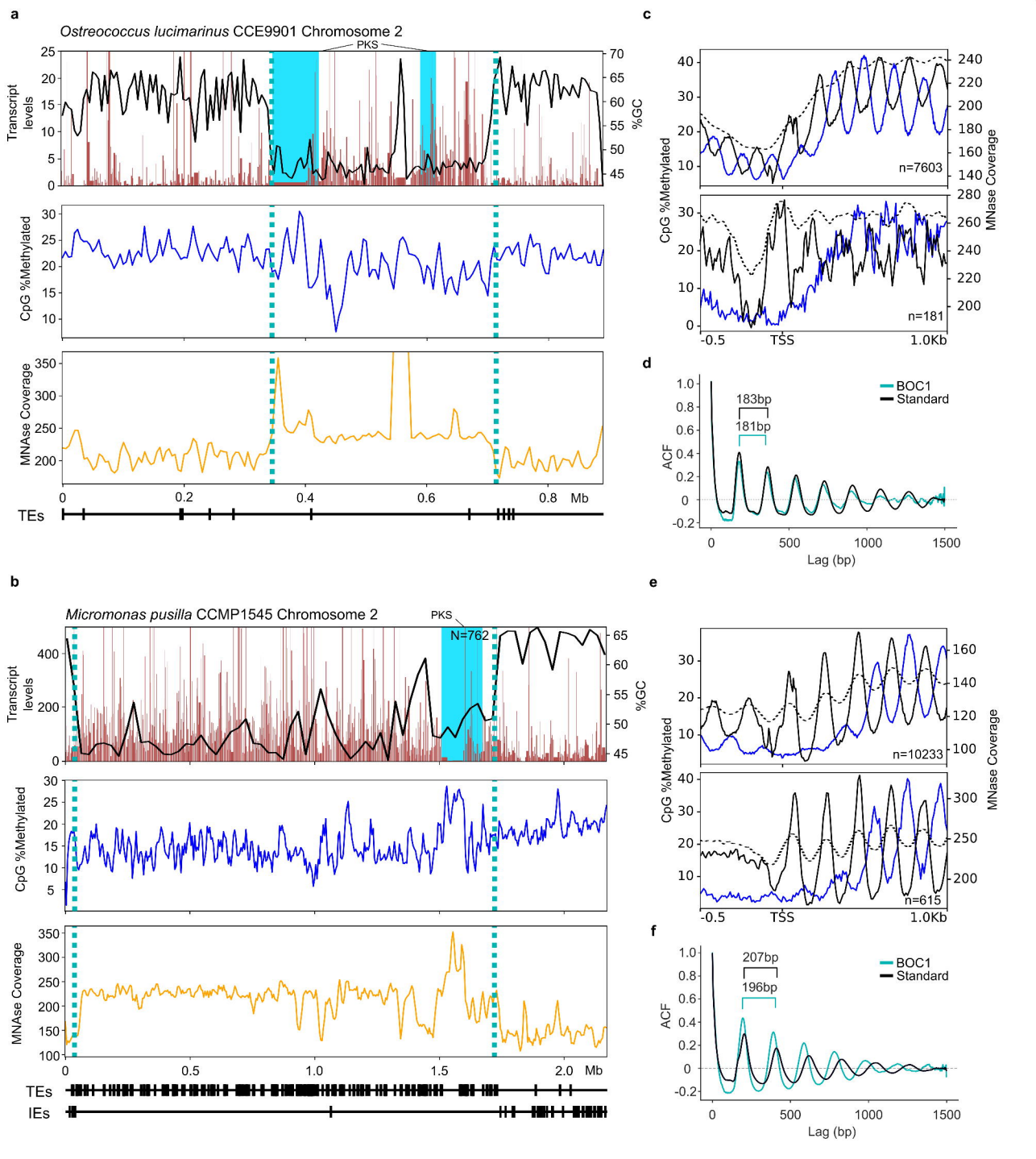
93. Krueger, F. & Andrews, S. R. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
94. Schultz, D. T. *et al.* Ancient gene linkages support ctenophores as sister to other animals. *Nature* **618**, 110–117 (2023).
95. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368 (2021).
96. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol* **20**, 275 (2019).
97. MacDougall, A. *et al.* UniRule: A unified rule resource for automatic annotation in the UniProt knowledgebase. *Bioinformatics* **36**, 4643–4648 (2020).
98. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **30**, 918–920 (2012).
99. Tsugawa, H. *et al.* MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods* **12**, 523–526 (2015).
100. Pang, Z. *et al.* MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. *Nucleic Acids Res* **52**, 398–406 (2024).
101. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods* **16**, 299–302 (2019).
102. Varadi, M. *et al.* AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **50**, 439–444 (2022).
103. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* **42**, 243–246 (2024).

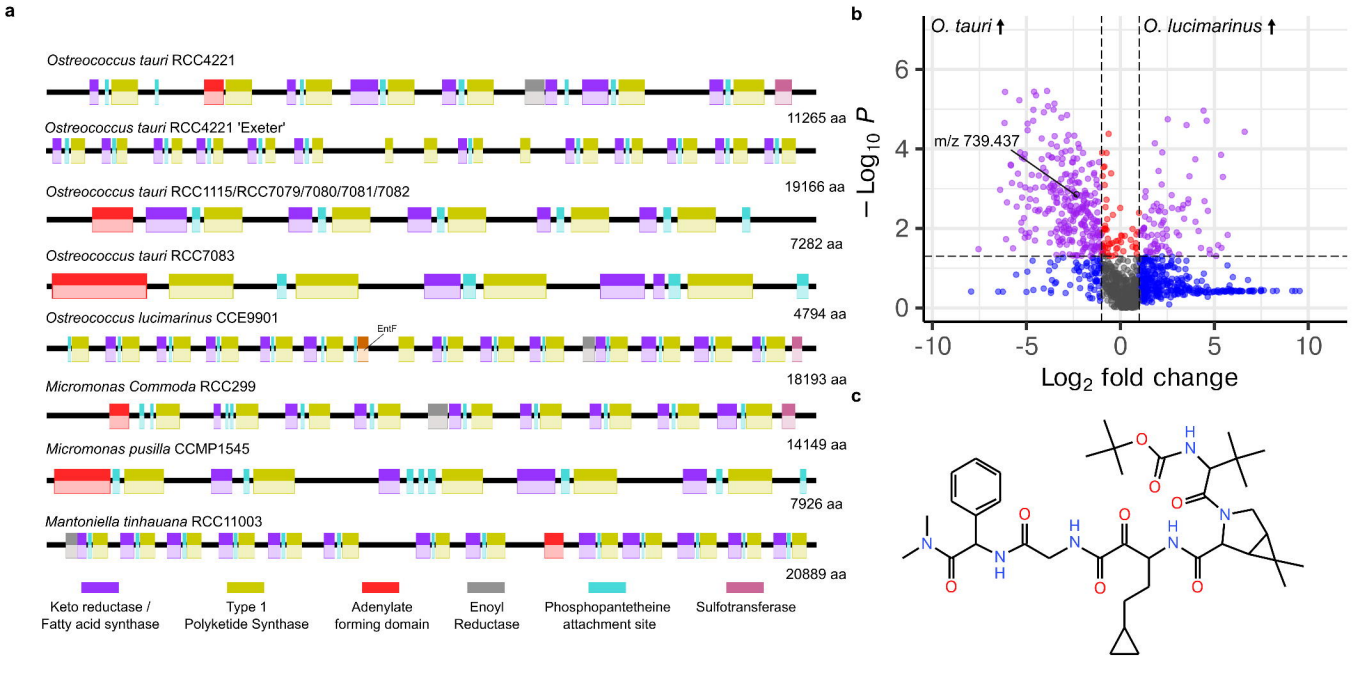

a
Ostreococcus tauri
Chromosome 2
1.07 Mb (~0.54 Mb)
Ostreococcus sp. RCC809
Chromosome 2
0.82 Mb (~0.32 Mb)
Ostreococcus lucimarinus
Chromosome 2
0.90 Mb (~0.36 Mb)
Bathycoccus prasinos
Chromosome 14
0.66 Mb (~0.39 Mb)
Micromonas commoda
Chromosome 1
2.05 Mb (~1.55 Mb)
Micromonas pusilla
Chromosome 2
2.17 Mb (~1.67 Mb)
b
Biological process
DNA transposition
DNA-templated transcription
DNA integration
glutaminyl-tRNAGln biosynthesis via transamidation
fatty acid biosynthetic process
photosynthesis
post-embryonic development
reproductive structure development
mitochondrial translation
histone deacetylation
SRP-dependent cotranslational protein targeting to membrane
Cellular component
nucleolus
transcription factor TFIIA complex
glutamyl-tRNA(Gln) amidotransferase complex
RNA polymerase II, core complex
proton-transporting V-type ATPase, V0 domain
COPI vesicle coat
MCM complex
proteasome core complex, alpha-subunit complex
retrograde vesicle-mediated transport, Golgi to endoplasmic reticulum
plastid
Molecular function
phosphopantetheine binding
DNA-directed 5'-3' RNA polymerase activity
3-oxoacyl-[acyl-carrier-protein] synthase activity
glutaminyl-tRNA synthase (glutamine-hydrolyzing) activity
chromatin binding
proton-transporting ATPase activity, rotational mechanism
ribosomal large subunit binding
DNA binding
hydrolase activity
histone deacetylase activity
translation initiation factor activity
ATP-dependent peptidase activity
2
3
4
Enrichment Value
Odds Ratio
10
20
30
40
50
gene count
3
6
9
12
15
18


a
Unknown
Retroelements
Non-LTR
Cre
LTR
Ty1/Copia
Ty3/Gypsy
DNA Transposons
DD(E/D) Transposons
Tc1/Mariner
PiggyBac
Harbinger
Mutator
hAT
Cacta
Rolling-circle
Helitron
YR Cryptons
Crypton-F
4221 St/B/S
1115 St/B/S
809 St/B/S
9901 St/B/S
1105 St/B/S
1545 St/B/S
299 St/B/S
11003 St/B/S
Ot
Ot
Os
Ol
Bp
Mp
Mc
Mt
TEs MB^{-1}
0
30
Mamiellales taxa
b
Standard
BOC1
SOC1
TE Density (TEs.MB^{-1})
0
100
200
300
400
500
600
700
Ot RCC4221
Ot RCC1115
Os
Ol
Bp
Mp
Mc
Mt
Mamiellales taxa
c Ty1/Copia
chr2:53,325-58,658
769
Copia
Hexosyltransferase*
(OT_ostta02g00190)
Ribonuclease
(OT_ostta02g00195)
d Tc1/Mariner
chr2:354,173-356,431
982
Transcription factor IIA
(OT_ostta02g01590)
Mariner
e Mutator
chr2:29,113-53,830
246
Mutator
Polyketide Synthase
(OT_ostta02g00160)
Polyketide Synthase
(OT_ostta02g00180)
f Mutator
chr2:186,000-224,354
401
unnamed product*
(OT_ostta02g00790)
Polyketide Synthase
(OT_ostta02g00800)
Mutator


a
Ostreococcus tauri Chromosome 2
N=532
Transcript levels
%GC
TEs
b
Micromonas commoda Chromosome 1
N=842
Transcript levels
%GC
TEs
c
Bathycoccus prasinos Chromosome 14
N=325
Transcript levels
%GC
TEs
d
p<2.2e-16
****
Transcript levels
Outlier
Standard
Region
e
p<2.2e-16
****
Transcript levels
Outlier
Standard
Region
f
p<2.2e-16
****
Transcript levels
Outlier
Standard
Region
g
Outlier
Standard
Transcript levels
%GC of nuclear CDS
h
Outlier
Standard
Transcript levels
%GC of nuclear CDS
i
Outlier
Standard
Transcript levels
%GC of nuclear CDS


a
Chromosome 2
(BOC)
Chromosome 19
(SOC)
250
Contact counts
0
b
Chromosome 2
1.07 Mb
TAD-like
350
Contact
counts
0
Transcript
levels
%GC
5mCG
5mC
6mA
5hmCG
5hmC
Avg. % of
Modifications
TEs
Chromosome position (Mb)
c
Chromosome 2
α
β
γ
δ
350
Contact
counts
0
Insulation
BOC1
Chromosome position (Mb)
d
Chromosome 19
300
Contact
counts
0
Insulation
SOC1
Chromosome position (Mb)
e
****
chr2/19
chr18/20
Interactions
Chr. 2 / 19
Other Chr.


a
Ostreococcus lucimarinus CCE9901 Chromosome 2
PKS
Transcript levels
%GC
CpG %Methylated
MNAse Coverage
TEs
Mb
b
Micromonas pusilla CCMP1545 Chromosome 2
PKS
N=762
Transcript levels
%GC
CpG %Methylated
MNAse Coverage
TEs
IEs
Mb
c
CpG %Methylated
MNase Coverage
n=7603
n=181
-0.5
TSS
1.0Kb
d
BOC1
Standard
183bp
181bp
ACF
Lag (bp)
e
CpG %Methylated
MNase Coverage
n=10233
n=615
-0.5
TSS
1.0Kb
f
BOC1
Standard
207bp
196bp
ACF
Lag (bp)


a
Ostreococcus tauri RCC4221
11265 aa
Ostreococcus tauri RCC4221 'Exeter'
19166 aa
Ostreococcus tauri RCC1115/RCC7079/7080/7081/7082
7282 aa
Ostreococcus tauri RCC7083
4794 aa
Ostreococcus lucimarinus CCE9901
EntF
18193 aa
Micromonas Commoda RCC299
14149 aa
Micromonas pusilla CCMP1545
7926 aa
Mantoniella tinhauana RCC11003
20889 aa
Keto reductase / Fatty acid synthase
Type 1 Polyketide Synthase
Adenylate forming domain
Enoyl Reductase
Phosphopantetheine attachment site
Sulfotransferase
b
O. tauri
O. lucimarinus
m/z 739.437
− Log10 P
Log2 fold change
c