Spatial Insulation Confines Dynamic Chromosomes in Minimal Marine Eukaryotes

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.

Corresponding author: Adam Monier; a.monier@exeter.ac.uk

Author Contributions: *Equal contributions

Competing Interest Statement: No competing interests

Keywords: Chromosome conformation capture, Genomic compartmentalisation, picoeukaryotes, transposable elements, polyketide synthase.

Abstract

Marine picoeukaryotes of the order Mamiellales, including *Ostreococcus tauri*, the smallest known free-living eukaryote, possess compact genomes yet maintain enigmatic "outlier chromosomes" characterised by lower GC content and hypervariability. To determine the structural and functional nature of these regions, we applied chromosome conformation capture to *O. tauri* and conducted comparative multi-omics analyses across the Mamiellales order, presenting the first analysis of three-dimensional genome organisation in marine picoeukaryotes. We reveal that outlier regions form structures resembling topologically associating domains, with sharp boundaries that spatially insulate them from the standard chromosomes. These compartments are defined by a distinctive chromatin state characterised by hypomethylation and transcriptional hyperactivity, and are frequently, though not universally, enriched in transposable elements. Crucially, species that lack transposable element enrichment in their outlier chromosomes nonetheless retain the transcriptional hyperactivity and distinct nucleotide composition of these regions, indicating that the functional identity of these compartments persists independently of transposon accumulation. The dynamic nature of these insulated domains is highlighted by the presence of structurally diverse giant polyketide synthase loci. We identify convergent genomic organisation in other chlorophytes, as well as phylogenetically distant stramenopiles. Our results suggest that such compartmentalisation of rapidly evolving, dynamic genomic regions represents a fundamental architectural principle of minimal eukaryotic genomes.

Introduction

The marine green alga *Ostreococcus tauri* is the smallest known free-living eukaryote (~0.8 µm), containing only a single plastid and mitochondrion (1, 2). This minimal cellular organisation presents a simplified model for studying fundamental eukaryotic processes such as photosynthesis, circadian rhythms, and environmental responses (3–5). *Ostreococcus* is a member of the Mamiellales (class Mamiellophyceae), a group that also includes other picoalgae such as *Bathycoccus*, *Mantoniella*, and *Micromonas* (6). Mamiellales are globally distributed, thriving in diverse marine environments from tropical to polar waters (7–9), supported by ecotype differentiation and intra-order genetic diversity (10–12).

A distinctive feature of Mamiellales genomes is the presence of two “outlier chromosomes”, known as the big and small outlier chromosomes (BOC and SOC), each containing regions of markedly lower GC content (referred to as BOC1 and SOC1) (13). These regions are enriched in introns, transposable elements (TEs), highly expressed genes (14–16), and they vary significantly among *Ostreococcus* strains, with polymorphisms associated with phenotypic differences in growth and viral susceptibility (17, 18). In *O. tauri*, the BOC contains a putative mating-type locus and genes linked to chromatin conformation (14, 17, 19), while the SOC contributes to antiviral defence (18, 20). Although outlier chromosomes have been recognised since the first Mamiellales genomes were sequenced (21),

fundamental questions about these unusual genomic regions remain unanswered. How are outlier chromosomes organised in three-dimensional nuclear space? What regulatory mechanisms distinguish them from the rest of the genome? And are their distinctive features conserved across the Mamiellales lineage, or specific to individual species?

We addressed these questions using a multi-omics approach–including chromosome conformation capture (Hi-C), transcriptomics, genome-wide methylation analysis, and metabolomics–to characterise the structural, epigenetic, and functional properties of outlier chromosomes across multiple Mamiellales species. Our DNA interaction analysis, the first in picoeukaryotes, reveals that BOC1 and SOC1 form spatially insulated compartments akin to topologically associating domains (TADs). These regions share conserved features across species: reduced DNA methylation, elevated gene expression, and enrichment of lineage-specific genes including giant type I polyketide synthases. Yet the underlying sequence composition varies markedly: TEs dominate in *Ostreococcus tauri* and *Micromonas* spp. but are largely absent from *O. lucimarinus* and *Bathycoccus prasinos*, which nonetheless retain the distinctive sequence and transcriptional profiles of these compartments. Strikingly, we find similar spatial insulation of equivalent outlier regions in the phylogenetically distant stramenopile *Pelagomonas calceolata*, suggesting this chromosomal organisation has arisen independently in unrelated lineages. Our findings reveal that conserved genomic compartmentalisation, rather than any single sequence-level mechanism, defines these unusual regions in minimal eukaryotic genomes.

Results

Outlier regions are conserved across Mamiellales and enriched in genes associated with genomic rearrangement

To determine whether BOC1 and SOC1 represent a conserved feature of Mamiellales genomes, we examined six species with high-quality genome assemblies. Despite their highly compact genomes (SI Appendix, Table S1 and S2), all species examined (*Ostreococcus* sp. RCC809, *O. tauri*, *O. lucimarinus*, *Micromonas commoda*, *M. pusilla*, and *Bathycoccus prasinos*) contain distinct low-GC outlier regions (14, 22–24). Even in *Mantoniella tinhauana*, where their presence had been uncertain (25), we identified two such regions with ~15% lower GC content than the surrounding genome (SI Appendix, Fig. S1 and Table S2). Although BOC1 and SOC1 regions vary considerably in size across lineages (Fig. 1a), their consistent presence suggests they are an ancient feature of Mamiellales. Comparative analysis revealed that BOC1 regions maintain greater evolutionary conservation than SOC1. While SOC1 regions showed limited synteny and fewer orthologues across genera (SI Appendix, Fig. S2), BOC1 regions shared more orthologues, though not strict synteny (Fig. 1a).

The gene content of these regions suggests a distinct functional profile. Gene Ontology enrichment analysis of *O. tauri* BOC1 revealed a high concentration of genes involved in DNA transposition, transcription, and integration (Fig. 1b; Dataset S1). The region also contains genes for chromatin remodelling, such as mini-chromosome maintenance (MCM) complex

components and sister chromatid cohesion protein 2 (SCC2). In contrast, SOC1 is enriched for genes related to RNA polymerase activity, membrane components, and methylation machinery (Dataset S2), indicating distinct functional profiles between the two outlier regions.

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

We conducted a survey of introns and TEs across Mamiellales genomes, confirming previous findings (21, 24–26) and revealing a distinct, non-random distribution of these features. In *O. tauri*, BOC1 and SOC1 contain 14.5-fold and 17.4-fold more TEs, respectively, than the standard genome (Kolmogorov–Smirnov test, $p<0.001$; Fig. 2a,b; SI Appendix, Fig. S3; Dataset S3). For example, Ty1/Copia retrotransposon density increases from 2.04 elements per Mb in the standard genome to 37.4 and 167 elements per Mb in BOC1 and SOC1, respectively. Similar patterns are observed across most Mamiellales species (Dataset S4).

We next examined the genomic context of TEs, as they 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; Dataset S5). 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 inserting upstream and downstream of coding segments (Fig. 2e,f).

Intron distribution shows a similar pattern: BOC1 regions consistently show higher intron density across taxa, for instance four-fold higher than the standard genome in *O. tauri* (Dataset S4).

Genes on outlier regions are more highly expressed and frequently spliced

Analysis of transcriptomic data from *O. tauri* and other Mamiellales species revealed that genes in BOC1 regions consistently show higher expression compared to those in the standard genome. In *O. tauri*, outlier genes show 3.8-fold higher mean expression (Kruskal-Wallis, $p<0.001$; Fig. 3a, d). Similar trends are observed in *M. commoda*, *M. pusilla*, and *B. prasinos* (2.4, 3.3, and 1.6-fold, respectively; $p<0.001$; Fig. 3a–f), although not all species display higher expression in SOC1 (Dataset S6). In most species, this elevated transcription is negatively correlated with GC content, such that lower GC content in outlier regions is associated with higher gene expression (Spearman's rank, $\rho=-0.37$, -0.29, and -0.49 for *O. tauri*, *M. commoda*, and *M. pusilla*, respectively; $p<0.001$; Fig. 3g–i; SI Appendix, Table S3). However, in *B. prasinos*, the trend is reversed, with higher GC content correlating with increased gene expression ($\rho=0.22$, $p<0.001$; Fig. 3i).

In addition to elevated expression, outlier regions contribute disproportionately to transcriptional diversity through alternative splicing. Although the BOC1 region in *O. tauri* contains far fewer genes than the standard genome, it generates 38% (1,216 of 3,215) of all

detected splice variants. Similar enrichment of splice variants is observed in *O. lucimarinus*, *B. prasinos*, and *M. commoda* (20%; Dataset S7).

Hi-C analysis reveals spatial insulation of outlier regions

To examine the three-dimensional organisation of outlier regions, we performed Hi-C on *O. tauri*. Contact maps (330 million contacts at 10 kb resolution; Fig. 4a; SI Appendix, Fig. S4) showed a simple architecture dominated by intrachromosomal interactions (77% of all contacts; Dataset S8). BOC and SOC displayed disproportionately high intrachromosomal contacts (~20% of total; Fig. 4; SI Appendix, Fig. S5).

BOC1 forms a clear topologically associating domain (TAD)-like structure (Fig. 4b, c; SI Appendix, Fig. S6), 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 (Dataset S9). SOC1 exhibits a similar but smaller insulated pattern (Fig. 4d). The Hi-C map of SOC suggests a TAD-like structure encompassing SOC1, although the small size of SOC1 limits the ability to computationally identify domains.

Quantitative analysis showed that BOC and SOC have significantly fewer interchromosomal contacts than standard regions (Fig. 4e; $p<0.0001$; Dataset S10). Instead, these outlier regions preferentially interact with each other, indicating physical separation from the rest of the genome. Analysis of *M. tinhauana* Hi-C data (25) revealed similar compartmentalisation patterns (SI Appendix, Fig. S1; Dataset S8), suggesting that spatial insulation of outlier regions is conserved across the Mamiellales lineage.

Outlier regions maintain distinctive epigenetic landscapes

Whole-genome methylation profiling revealed that epigenetic and spatial boundaries are precisely aligned. 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. Methylation levels within outlier regions are 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; SI Appendix, Fig. S7 and S8; Dataset S11). 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 (SI Appendix, Fig. S9).

Available bisulfite sequencing data (27) 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; SI Appendix, Fig. S10; Dataset S12).

TE density correlates with altered nucleosome spacing and chromatin flexibility

157 The variation in TE abundance between Mamiellales species provides an opportunity to
158 examine whether TE load correlates with chromatin structure. We re-analysed published
159 MNase-seq data from *O. lucimarinus* and *M. pusilla* (27) to determine how variation in TE
160 density corresponds to nucleosome spacing and positioning within BOC1. Across both
161 species, nucleosome organisation in BOC1 bears hallmarks of transcriptionally permissive
162 chromatin: high nucleosome occupancy around transcription start sites, weaker periodicity
163 in promoters, and reduced CpG methylation (Fig. 5c,e). These features coincide with
164 elevated transcription: RNA-Seq coverage in BOC1 is 1.86× and 2.44× higher than in standard
165 regions for *O. lucimarinus* and *M. pusilla*, respectively (Kruskal–Wallis, $p < 0.001$).

166 Nucleosome spacing differs markedly between species and genomic compartments. In the
167 TE-rich *M. pusilla*, nucleosome repeat length shortens from 207 bp in standard regions to 196
168 bp in BOC1, whereas the TE-poor *O. lucimarinus* shows only a minor shift from 183 bp to 181
169 bp (Fig. 5d,f). In both species, outlier regions show reduced chromatin stiffness and more
170 well-positioned nucleosomes. Comparisons with invariant chromosomes further support
171 these findings (SI Appendix, Fig. S11; Dataset S13).

172 TE–nucleosome spatial relationships further distinguish these regions. Whereas *O.*
173 *lucimarinus* BOC1 shows little TE enrichment, *M. pusilla* BOC1 contains 3.5× more TEs than
174 its standard genome. Its BOC1 also shows a depletion in introner elements (IEs), which are
175 otherwise abundant outside this region (24, 28). TEs display a bimodal distribution with
176 peaks ±50 bp from nucleosome centres, especially pronounced in BOC1, consistent with
177 preferential insertion into accessible linker DNA rather than nucleosome cores (SI Appendix,
178 Fig. S12). Notably, the same regions encoding giant type I PKS genes exhibit elevated MNase
179 coverage and CpG methylation (Fig. 5a,b).

180 **Giant type I polyketide synthase genes show structural diversity with distinct**
181 **metabolic outputs**

182 Giant type I PKS genes, a conserved feature of BOC1 regions, show large structural variations
183 across Mamiellales species. PKS gene architecture varies dramatically, with lengths ranging
184 from 14,393 bp in *O. tauri* to 64,971 bp in *M. tinhauana* (Fig. 6a; Dataset S14). Even within
185 species, intraspecific variation is substantial: different *O. tauri* strains harbour PKS genes
186 ranging from 14,393 to 33,947 bp. Long-read resequencing of our *O. tauri* RCC4221
187 laboratory culture ("Exeter") revealed that the overall BOC1 architecture remains conserved
188 relative to the reference genome, but the PKS locus displays substantial structural
189 reorganisation, with additional modules, domain rearrangements, and multiple inversions
190 (Fig. 6a; Dataset S14).

191 To explore whether these structural variations correspond to metabolic divergence, we
192 performed untargeted metabolomics on *O. tauri* (RCC4221) and *O. lucimarinus* (CCE9901),
193 two species with contrasting PKS organisations in BOC1 (SI Appendix, Fig. S13; Dataset S15).
194 Across both ionisation modes, we detected 2,213 unknown mass features, with 273 and 50
195 showing significantly higher abundance in *O. tauri* and 81 and 41 in *O. lucimarinus* in positive
196 and negative ion modes, respectively (ANOVA, $p<0.05$, $|Log2FC|>2$; Fig. 6b; Dataset S16 and
197 S17). These comparisons reveal broad chemical divergence between the species. Among the
198 discriminant features, we identified one compound (RT 14.38 min, m/z 739.437; Fig. 6c; SI

199 Appendix, Fig. S14; Dataset S18) whose fragmentation pattern is consistent with polyketide-
200 like structures, though structural confirmation and direct demonstration of PKS involvement
201 remain to be established.

202 **Similar outlier architecture in phylogenetically distant marine picoeukaryotes**

203 To determine whether spatially insulated outlier regions occur in other minimal marine
204 eukaryotes, we analysed genomic data from diverse lineages, including those sequenced by
205 the Darwin Tree of Life project (29). In the stramenopile *Pelagomonas calceolata* (30), we
206 identified discrete ~500 kb outlier regions on each chromosome characterised by ~5% lower
207 GC content and significantly elevated TE density (Mann-Whitney U test $p<0.0001$; SI
208 Appendix, Fig. S15). Analysis of available Hi-C data confirmed that these regions are spatially
209 compartmentalised from the standard genome (30), exhibiting insulation architecture
210 similar to that observed in Mamiellales BOC1 and SOC1. Genes within *P. calceolata* outlier
211 regions show ~3-fold elevated expression compared to standard regions. Remarkably, one
212 *P. calceolata* outlier region contains a giant PKS gene positioned adjacent to a MULE element
213 (SI Appendix, Fig. S16). We also identified similar outlier features in the diatom *Epithemia*
214 *pelagica* (Bacillariophyceae) (31), where specific chromosomes (e.g., OX337238.1) harbour
215 dense, localised clusters of TEs, predominantly LTR retrotransposons of the Gypsy
216 superfamily (97 copies in the outlier region; Dataset S19).

217 We next examined other Chlorophyta classes. In the Pseudoscourfieldiophyceae, we
218 analysed the genomes of *Pycnococcus provasolii* and *Pseudoscourfieldia marina* (32).
219 Although historically classified as distinct genera, recent organellar phylogenomics has
220 unified them as morphotypes of the single species *P. marina* (32). Consistent with this
221 reclassification, we observed 98% average nucleotide identity between the two assemblies
222 and high synteny across their standard chromosomes. However, we identified distinct, low-
223 GC outlier regions in both genomes (Mann-Whitney U test, $p<0.0001$; Dataset S19). While the
224 standard chromosomes are highly conserved, the outlier regions exhibit marked structural
225 divergence and TE load differences between the two morphotypes (SI Appendix, Fig. S17),
226 paralleling the intraspecific divergence observed in *Ostreococcus*. Finally, in the
227 trebouxiophyte *Marvania coccoides* (33), we identified a single chromosome (OY978099.1)
228 containing a discrete region with massive TE accumulation (362 elements, including 106 LTR
229 retrotransposons, compared to a genome average of <50 per comparable window; Mann-
230 Whitney U test $p<0.0001$; Dataset S19).

231 The recurrence of spatially defined, TE-rich, low-GC compartments across the chlorophytes
232 (Mamiellophyceae, Pseudoscourfieldiophyceae, and Trebouxiophyceae), and
233 stramenopiles (Pelagophyceae and Bacillariophycea), indicates that this chromosomal
234 organisation has arisen independently in phylogenetically distant lineages of marine
235 eukaryotes.

236 Discussion

237 **Outlier regions form spatially insulated compartments with distinctive chromatin**
238 **properties**

239 Mamiellales, among the smallest free-living eukaryotes, harbour unusually large and
240 dynamic chromosomal regions (BOC1 and SOC1) that represent distinct genomic
241 compartments. Our multi-omics analysis reveals that these outlier regions share a
242 conserved suite of structural features across species: reduced DNA methylation, elevated
243 gene expression, and spatial insulation from the rest of the genome. The Hi-C data presented
244 here demonstrate that BOC1 and SOC1 form TAD-like compartments, with sharp boundaries
245 that align precisely with epigenetic transitions and changes in TE density.

246 The transcriptional hyperactivity of genes within outlier regions mirrors patterns observed in
247 low-GC regions of land plants such as *Oryza sativa* (34), but contrasts with the freshwater
248 alga *Chlamydomonas reinhardtii*, where low GC content correlates with reduced expression
249 (35). This contrast suggests that marine picoeukaryotes and their freshwater relatives have
250 evolved distinct regulatory architectures governing gene expression in low-GC regions. The
251 hypomethylated state of outlier regions, combined with their spatial insulation, is associated
252 with a chromatin environment that is transcriptionally permissive and physically separated
253 from the compact, gene-dense standard genome.

254 **Nucleosome organisation is structured and may influence TE distribution**

255 Contrary to earlier cryo-electron tomography studies suggesting disordered chromatin in *O.*
256 *tauri* (36, 37), our re-analysis of MNase-seq data (27) reveals that chromatin organisation in
257 these minimal eukaryotes is structured and species-specific. The outlier regions in *O.*
258 *lucimarinus* and *M. pusilla* display well-phased, closely spaced nucleosomes, which is a
259 configuration associated with chromatin flexibility (38, 39). In *M. pusilla*, the bimodal
260 distribution of TEs around nucleosome centres suggests that chromatin structure may
261 influence TE insertion patterns, with elements preferentially integrating into accessible linker
262 DNA. This parallels findings in *Arabidopsis*, where the histone variant H2A.Z influences TE
263 insertion site preference (40). Given the presence of plant-like H2A.Z histone variants, which
264 are known to influence chromatin structure and TE insertion, in Mamiellales, similar
265 mechanisms may contribute to the concentration of TEs within outlier regions.

266 **Three-dimensional organisation, not TE content, defines outlier regions**

267 A key finding of our study is that several features of outlier regions are conserved across
268 Mamiellales, regardless of TE content. This is most clearly illustrated by *O. lucimarinus* and
269 *B. prasinos*, which lacks substantial TE enrichment yet retains the characteristic low-GC
270 profile and elevated gene expression observed in TE-rich species. While Hi-C data is
271 unavailable for both TE-depleted species, the persistence of these hallmark features
272 suggests that the underlying genomic context, likely involving conserved chromatin
273 regulation, is maintained independently of TE accumulation. This argues against models in
274 which TEs are the primary drivers of outlier region properties, instead suggesting that TEs
275 accumulate as a consequence of the unique genomic environment of these regions.

276 **Giant PKS genes exhibit extensive structural variation**

277 The giant type I PKS genes found in BOC1 represent one of the most striking examples of
278 structural variation within these regions. PKS enzymes synthesise diverse secondary

metabolites including toxins, antibiotics, and signalling molecules that can be important for ecological interactions (41, 42). We found that MULEs frequently flank or insert within PKS loci. Since MULEs are known to mediate gene duplication and shuffling (43), their proximity to PKS genes is notable, though whether they actively contribute to PKS diversification remains to be determined. Our long-read sequencing revealed substantial structural variation within the PKS locus of *O. tauri* RCC4221 compared to the reference genome, and untargeted metabolomics detected species-specific chemical differences between *O. tauri* and *O. lucimarinus*. Among these, one feature showed properties compatible with a polyketide-like compound, though its structure and biosynthetic origin remain unconfirmed. Establishing a direct link between PKS gene variation and metabolic output will require targeted biochemical characterisation.

Structural parallels with sex-determining regions

The structural features of BOC1, suppressed recombination and TE accumulation, resemble the sex-determining regions of other eukaryotes (44, 45). This organisation parallels the UV sex chromosome systems of brown algae such as *Ectocarpus*, where non-recombining regions accumulate TEs and lineage-specific genes (46, 47). Recent Hi-C analysis of *Ectocarpus* confirms that its sex-determining regions are spatially insulated from flanking pseudoautosomal regions (48), directly paralleling the TAD-like compartmentalisation we observe in Mamiellales. However, an important distinction exists: while *Ectocarpus* pseudoautosomal regions often show lower expression (47), the *Ostreococcus* outlier regions are transcriptionally hyperactive. Whether the outlier regions of Mamiellales evolved from ancestral mating-type loci, or whether the similarities reflect convergent structural solutions to distinct biological problems, remains an open question.

Convergent genome compartmentalisation across marine picoeukaryotes

The identification of similar outlier architecture in phylogenetically distant lineages is perhaps the most unexpected finding of our comparative analysis. In the stramenopile *P. calceolata*, we found discrete low-GC regions with elevated TE density, spatial insulation (30), increased gene expression, and, remarkably, giant PKS genes adjacent to MULE elements. Similar features occur in the diatom *E. pelagica* (31) and in chlorophyte lineages outside the Mamiellophyceae. The recurrence of this chromosomal organisation across Mamiellophyceae, Pseudoscourfieldiophyceae, Trebouxiophyceae, Pelagophyceae, and Bacillariophyceae suggests either deep ancestry predating the divergence of these groups, or, more likely given the phylogenetic distances involved, independent origins of similar genome organisation. What selective pressures or mechanistic constraints might favour this arrangement remains unclear. One possibility is that spatial insulation provides a means to tolerate TE activity without disrupting essential genes; another is that the chromatin environment of these regions is permissive for the expression of large, complex genes like PKS loci that might be poorly expressed elsewhere in compact genomes. These hypotheses are not mutually exclusive and remain to be tested.

Tolerance without silencing: an unconventional epigenetic architecture

319 Our findings contribute to an emerging picture of how genome architecture varies across
320 diverse eukaryotes. Similar patterns of compartmentalised genomic divergence have
321 recently been reported in bivalve mollusks, where low-GC, TE-rich "Haplotype Divergent
322 Sequences" show elevated structural variation (49). In contrast to many eukaryotes, that
323 silence repetitive regions via hypermethylation (50, 51), the outlier regions of minimal marine
324 eukaryotes maintain a hypomethylated, transcriptionally active state. Whether this reflects
325 distinct regulatory mechanisms, weaker silencing machinery, or active maintenance of an
326 "open" chromatin state remains to be determined. Key questions for future work include the
327 molecular mechanisms that establish and maintain the sharp boundaries between outlier
328 and standard regions, whether outlier chromosome polymorphisms contribute to
329 phenotypic traits such as viral susceptibility (17, 18), and the evolutionary history of these
330 regions across eukaryotic lineages. Our analysis establishes that outlier chromosomes are
331 spatially insulated compartments defined primarily by their three-dimensional organisation,
332 rather than by specific sequence composition. The independent occurrence of similar
333 chromosomal architecture in distantly related marine eukaryotes points to shared
334 fundamental principles, whether adaptive or mechanistic, that govern genome organisation
335 in compact eukaryotic genomes.

336 Materials and Methods

337 Algal strains and culturing

338 A total of eight Mamiellales species were included in the study, representing the genera
339 *Ostreococcus*, *Bathycoccus*, *Micromonas* and *Mantoniella*. We used *Ostreococcus tauri*
340 RCC4221 in this study to generate data for Hi-C, RNA-Seq and methylation profiling. We
341 sourced other molecular datasets from publicly available databases. The species, strain IDs
342 and datasets used in this study are provided in SI Appendix, Table S1.

343 *O. tauri* strain RCC4221 was maintained in L1 seawater medium (52) in non-treated
344 polycarbonate flasks and maintained in a 14 h: 10 h light: dark cycle at 35 μmol m^{-2} s^{-1}
345 irradiance and a temperature of 20 °C. To achieve a near-axenic culture of *O. tauri*, we
346 isolated cells by flow cytometric cell sorting (BD Influx, Beckton Dickinson) and maintained
347 algal cells in exponential growth by weekly subculturing. We also used a monthly antibiotic
348 treatment (5 μg Ampicillin, 1 μg Gentamicin, 2 μg Kanamycin, 10 μg Neomycin). Algal growth
349 was regularly monitored by flow cytometry (Cytoflex, Beckman Coulter) based on side-angle
350 light scatter and chlorophyll red fluorescence. Bacterial cells were monitored and quantified
351 via flow cytometry after SYBR Green I staining.

352 Transcriptomics

353 **RNA extraction and sequencing.** *O. tauri* RCC4221 cells were collected by centrifuging a 20
354 mL sample (10 min at 4000 × g at 4 °C), snap-frozen in liquid nitrogen, and stored at −80 °C.
355 RNA was extracted using a RNeasy Plant kit (QIAGEN) with RLT buffer incubated at 56 °C for
356 3 min. The residual DNA was digested using the Turbo DNA-free kit (Thermo Fisher Scientific).
357 RNA was quantified using a Qubit Fluorometer (Thermo Fisher Scientific), and RNA quality
358 was assessed using an Agilent 2100 Bioanalyzer System. RNA extracts with an RNA integrity
359 number (RIN) greater than eight were subjected to poly(A) selection prior to library

360 preparation with the Illumina TruSeq Direction Library Prep Kit and sequenced on Illumina
361 NovaSeq 6000 (150 bp paired end reads).

362 **RNA-Seq data processing.** Adapters and low-quality reads from RNA-Seq libraries were
363 trimmed using Fastp v0.23.2 (53) (minimum length required 50). Quality control reports were
364 compiled using FastQC v0.11.9 (54), MultiQC v1.13 (55), and fastp. Ribosomal RNA reads
365 were removed using SortMERNA v4.3.6 (56) matching reads with rRNA sequences from the
366 SILVA databases (57). The remaining reads were aligned and quantified to the *O. tauri*
367 reference genome sequence (GCF_000214015.3) using Star v2.7.10 (58) with default
368 parameters. Counts were divided by gene length to standardise counts and generate
369 transcript levels. For publicly available RNA-Seq libraries and counts, transcript levels were
370 measured in *M. pusilla* CCMP1545 using table 'Sup_Table_S2_CCMP1545' column 'P-
371 Replete T1.5 hours' for counts (59). For *M. commoda* RCC299, existing triplicate axenic
372 counts were used (60). For *B. prasinos* RCC1105, *O. lucimarinus* CCE9901, *M. commoda*
373 RCC299, the reads from SRR1300453 (61), SRR847308 (27) and SRR17220685 (60) were used
374 with the same workflow as above to additionally measure splice events with Star.

375 **Oxford Nanopore Technology sequencing**

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

383 **Methylation profiling.** Super-accuracy base calling was performed on R10.4.1 sequence
384 data using Dorado v0.6.0 (ONT, Oxford, UK), Methylation profiling included N6-
385 methyladenine (6mA), N5-methylcytosine (5mC/5mCG), and N5-hydroxymethylcytosine
386 (5hmC/5hmCG) which were transformed into BED files using modkit v0.2.8 (ONT, Oxford,
387 UK). To normalise for variations in local GC content, % methylation levels represent the
388 average proportion of the methylated base relative to the total number of that base within the
389 sliding window.

390 **Genome assembly.** Exeter RCC4221 was de novo assembled from R9.4.1 ONT sequence
391 data using Flye v2.9.2 (62) (--iterations 3 –scaffold) followed by medaka_consensus v1.8.0
392 (ONT, Oxford, UK). Matching scaffolds to reference were pulled out and aligned using blastn
393 v2.1.0 (63).

394 **High-throughput chromosome conformation capture**

395 **Hi-C sequencing.** For consistency, samples for all Hi-C experiments were collected at 2
396 hours into the light phase. To ensure sufficient DNA yield for sequencing from small *O. tauri*
397 genomes, 1–3 × 10^8 total cells were used in all experiments. As there was no published
398 protocol for picoeukaryotes we tested both *in situ* (64) and *in vivo* Hi-C approaches and a
399 series of optimisation steps, which are detailed in the SI Appendix. 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 Dataset S10).

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 (54), MultiQC v1.13 (55), 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) (65–67). Hi-C reads were mapped onto a soft and hard-masked (SI Appendix, Fig. S18) *O. tauri* reference genome sequence (GenBank identifier: GCF_000214015.3) using Bowtie v2.3.5.1 (68). Duplication removal and contact pair validation were conducted with HiC-Pro v2.11.1 (65). Raw and normalised contact maps were generated using cooltools v0.3.0 (69) 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 (70). Hi-C quality reports were generated using HiCUP (71). Domain compartment calling was completed using Cooltools. The R package HiCRep (72) was used to evaluate the correlation between Hi-C experimental replicates (SI Appendix, Fig. S19). Juicer v1.6 (73) 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 (74) and HiCCUPS v3.0 (73, 74), respectively, at bin resolutions of 5, 10, and 20 kb. The derivative/slope of contact probability, cis, and trans interactions was calculated using cooltools. *M. tinhauana* RCC11003 Hi-C data was similarly analysed (25).

Additional sequence analyses

MNase-seq analysis. MNase-seq data for *M. pusilla* CCMP1545 and *O. lucimarinus* CCE9901 (SRR847304 and SRR847307) (27) 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 (75) using default parameters for paired-end data. Aligned files were converted to BAM format, sorted, and indexed with Samtools v1.13 (76). Finally, genome-wide coverage was calculated and normalised to RPKM using bamCoverage (deeptools v3.5.1 (77), --MNase --center-reads) to calculate nucleosome centres, 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 (78).

Bisulfite-seq analysis. Bisulfite sequencing data for *M. pusilla* CCMP1545 and *O. lucimarinus* CCE9901 (SRR847303 and SRR847306 (27)) 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 (79). 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 (80), blastp v2.10 (63) and Diamond v2.1.8 (81) across outlier chromosomes in the Mamiellophyceae. A ribbon plot was generated to show the synteny between species using Python. Updated protein annotations were determined using AlphaFold3 (82) and FoldSeek (83). EDTA v2.2.2 (84), 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 (85). 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 with 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 20,000 × g 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 particle size reverse phase C18 analytical column (Agilent Technologies, Palo Alto, USA). 5 µL samples were injected *via* an Agilent 1200 series Rapid Resolution HPLC system coupled to a quadrupole time-of-flight (QToF) 6520 mass spectrometer (Agilent Technologies, Palo Alto, USA). Mobile phases were A (10% methanol, 0.1% formic acid), mobile phase B (100% methanol with 0.1% formic acid). The following gradient was used: 0–5 min-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 was 40 °C. The source conditions for electrospray ionisation were as follows: gas temperature was 325 °C with 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 positive and negative ion mode. The fragmentor voltage was 175 V and skimmer 65 V. Scanning was performed 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 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 compound m/z values were 121.05087 and 922.00979. Raw data files (*.d) were converted to mzML format using MSConvert (ProteoWizard) (86). Feature detection and alignment was carried out with MS-Dial v5 (87). Features present in the blank (>20% of sample intensity) were discounted. Compounds were identified with the MS-Dial lipidomics workflow and using publicly available CompMS MS-Dial MS/MS libraries. Statistical one-factor Anova test analysis ($p<0.05$, $|LFC|>2$) was then performed on 'Unknown' features with Metaboanalyst v6.0 (88) to highlight putative PKS products for further annotation within Sirius (89).

Data availability

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

Acknowledgements

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

Supporting Information

Supporting Information (PDF)
Dataset_Supporting (XLSX)

References

1. C. Courties, *et al.*, Smallest eukaryotic organism. *Nature* 370, 255 (1994).

2. M. J. Chretiennot-Dinet, *et al.*, A new marine picoeucaryote: Ostreococcus tauri gen. et sp. nov. (Chlorophyta, Prasinophyceae). *Phycologia* 34, 285–292 (1995).

3. Z. B. Noordally, *et al.*, A phospho-dawn of protein modification anticipates light onset in the picoeukaryote Ostreococcus tauri. *J. Exp. Bot.* 74, 5514–5531 (2023).

4. F. Corellou, *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. S. F. Martin, *et al.*, Surviving starvation: Proteomic and lipidomic profiling of nutrient deprivation in the smallest known free-living eukaryote. *Metabolites* 10, 273 (2020).

6. B. Marin, M. Melkonian, 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. M. Tragin, D. Vaulot, Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding. *Sci. Rep.* 9, 5190 (2019).

8. C. C. M. Yung, E. R. Redondo, F. Sanchez, S. Yau, G. Piganeau, Diversity and Evolution of Mamiellophyceae: Early-Diverging Phytoplanktonic Green Algae Containing Many Cosmopolitan Species. *J. Mar. Sci. Eng.* 10, 240 (2022).

9. D. Demory, *et al.*, Picoeukaryotes of the Micromonas genus: sentinels of a warming ocean. *ISME Journal* 13, 132–146 (2019).

10. E. Demir-Hilton, *et al.*, Global distribution patterns of distinct clades of the photosynthetic picoeukaryote Ostreococcus. *ISME Journal* 5, 1095–1107 (2011).

523 11. A. Monier, A. Z. Worden, T. A. Richards, Phylogenetic diversity and biogeography of the
524 Mamiellophyceae lineage of eukaryotic phytoplankton across the oceans. *Environ.*
525 *Microbiol. Rep.* 8, 461–469 (2016).

526 12. E. Sands, *et al.*, Genetic and physiological responses to light quality in a deep ocean
527 ecotype of Ostreococcus, an ecologically important photosynthetic picoeukaryote. *J. Exp.*
528 *Bot.* 74, 6773–6789 (2023).

529 13. N. Grimsley, S. Yau, G. Piganeau, H. Moreau, "Typical features of genomes in the
530 mamiellophyceae" in *Marine Protists: Diversity and Dynamics*, (Springer, 2015), pp. 107–
531 127.

532 14. H. Moreau, *et al.*, Gene functionalities and genome structure in Bathycoccus prasinos
533 reflect cellular specializations at the base of the green lineage. *Genome Biol.* 13, R74
534 (2012).

535 15. W. Lanier, A. Moustafa, D. Bhattacharya, J. M. Comeron, EST analysis of Ostreococcus
536 lucimarinus, the most compact eukaryotic genome, shows an excess of introns in highly
537 expressed genes. *PLoS One* 3, e2171 (2008).

538 16. M. Irimia, S. W. Roy, Evolutionary convergence on highly-conserved 3′ intron structures in
539 intron-poor eukaryotes and insights into the ancestral eukaryotic genome. *PLoS Genet.* 4,
540 e1000148 (2008).

541 17. R. Blanc-Mathieu, *et al.*, Population genomics of picophytoplankton unveils novel
542 chromosome hypervariability. *Sci. Adv.* 3, e1700239 (2017).

543 18. S. Yau, *et al.*, A Viral Immunity Chromosome in the Marine Picoeukaryote, Ostreococcus
544 tauri. *PLoS Pathog.* 12, e1005965 (2016).

545 19. L. F. Benites, *et al.*, Evolutionary Genomics of Sex-Related Chromosomes at the Base of
546 the Green Lineage. *Genome Biol. Evol.* 13, evab216 (2021).

547 20. S. Yau, *et al.*, Virus-host coexistence in phytoplankton through the genomic lens. *Sci. Adv.*
548 6, eaay2587 (2020).

549 21. E. Derelle, *et al.*, Genome analysis of the smallest free-living eukaryote Ostreococcus
550 tauri unveils many unique features. *Proc. Natl. Acad. Sci. U. S. A.* 103, 11647–11652
551 (2006).

552 22. R. Blanc-Mathieu, *et al.*, An improved genome of the model marine alga Ostreococcus
553 tauri unfolds by assessing Illumina de novo assemblies. *BMC Genomics* 15, 1103 (2014).

554 23. B. Palenik, *et al.*, The tiny eukaryote Ostreococcus provides genomic insights into the
555 paradox of plankton speciation. *Proc. Natl. Acad. Sci. U. S. A.* 104, 7705–7710 (2007).

556 24. A. Z. Worden, *et al.*, Green evolution and dynamic adaptations revealed by genomes of
557 the marine picoeukaryotes micromonas. *Science (1979).* 324, 268–272 (2009).

558 25. E. Rey Redondo, Y. Xu, C. C. M. Yung, Genomic characterisation and ecological
559 distribution of Mantoniella tinhauana: a novel Mamiellophycean green alga from the
560 Western Pacific. *Front. Microbiol.* 15, 1358574 (2024).

561 26. S. Jancek, S. Gourbière, H. Moreau, G. Piganeau, Clues about the genetic basis of
562 adaptation emerge from comparing the proteomes of two Ostreococcus ecotypes
563 (Chlorophyta, Prasinophyceae). *Mol. Biol. Evol.* 25, 2293–2300 (2008).

564 27. J. T. Huff, D. Zilberman, Dnmt1-independent CG methylation contributes to nucleosome
565 positioning in diverse eukaryotes. *Cell* 156, 1286–1297 (2014).

566 28. M. P. Simmons, *et al.*, Intron invasions trace algal speciation and reveal nearly identical
567 arctic and antarctic micromonas populations. *Mol. Biol. Evol.* 32, 2219–2235 (2015).

568 29. M. L. Blaxter, Sequence locally, think globally: The Darwin Tree of Life Project. *Proc. Natl.*
569 *Acad. Sci. U. S. A.* 119, e2115642118 (2022).

570 30. N. Guérin, *et al.*, Genomic adaptation of the picoeukaryote Pelagomonas calceolata to
571 iron-poor oceans revealed by a chromosome-scale genome sequence. *Commun. Biol.* 5,
572 983 (2022).

573 31. C. R. Schvarcz, *et al.*, The genome sequences of the marine diatom Epithemia pelagica
574 strain UHM3201 (Schvarcz, Stancheva & Steward, 2022) and its nitrogen-fixing,
575 endosymbiotic cyanobacterium. *Wellcome Open Res.* 9, 232 (2024).

576 32. O. Crépeault, C. Otis, J. F. Pombert, M. Turmel, C. Lemieux, Comparative plastome and
577 mitogenome analyses indicate that the marine prasinophyte green algae Pycnococcus
578 provasolii and Pseudoscourfieldia marina (Pseudoscourfieldiophyceae class nov.,
579 Chlorophyta) represent morphotypes of the same species. *J. Phycol.* 60, 1021–1027
580 (2024).

581 33. R. Allen, *et al.*, The genome sequence of the chlorophyte Marvania coccoides CCAP
582 251/1B (Naumann) Henley, Hironaka, Guillou, M. Buchheim, J. Buchheim, M. Fawley & K.
583 Fawley 2004. *Wellcome Open Res.* 9, 334 (2024).

584 34. S. Das, M. Bansal, Variation of gene expression in plants is influenced by gene
585 architecture and structural properties of promoters. *PLoS One* 14, e0212678 (2019).

586 35. R. Barahimipour, *et al.*, Dissecting the contributions of GC content and codon usage to
587 gene expression in the model alga Chlamydomonas reinhardtii. *Plant Journal* 84, 704–717
588 (2015).

589 36. L. Gan, M. S. Ladinsky, G. J. Jensen, Chromatin in a marine picoeukaryote is a
590 disordered assemblage of nucleosomes. *Chromosoma* 122, 377–386 (2013).

591 37. L. Gan, M. S. Ladinsky, G. J. Jensen, Organization of the smallest eukaryotic spindle.
592 *Current Biology* 21, 1578–1583 (2011).

593 38. H. Ohtani, Y. W. Iwasaki, Rewiring of chromatin state and gene expression by
594 transposable elements. *Dev. Growth Differ.* 63, 262–273 (2021).

595 39. A. Huda, L. Mariño-Ramírez, D. Landsman, I. K. Jordan, Repetitive DNA elements,
596 nucleosome binding and human gene expression. *Gene* 436, 12–22 (2009).

597 40. L. Quadrana, *et al.*, Transposition favors the generation of large effect mutations that may
598 facilitate rapid adaption. *Nat. Commun.* 10, 3421 (2019).

599 41. A. Nivina, K. P. Yuet, J. Hsu, C. Khosla, Evolution and Diversity of Assembly-Line
600 Polyketide Synthases. *Chem. Rev.* 119, 12524–12547 (2019).

601 42. T. R. Fallon, *et al.*, Giant polyketide synthase enzymes in the biosynthesis of giant marine
602 polyether toxins. *Science* 385, 671–678 (2024).

603 43. M. Dupeyron, K. S. Singh, C. Bass, A. Hayward, Evolution of Mutator transposable
604 elements across eukaryotic diversity. *Mob. DNA* 10, 12 (2019).

605 44. N. Grimsley, B. Péquin, C. Bachy, H. Moreau, G. Piganeau, Cryptic sex in the smallest
606 eukaryotic marine green alga. *Mol. Biol. Evol.* 27, 47–54 (2010).

607 45. D. Charlesworth, Evolution of recombination rates between sex chromosomes.
608 *Philosophical Transactions of the Royal Society B: Biological Sciences* 372, 20160456
609 (2017).

610 46. E. Dinatale, R. J. Craig, C. Martinho, H. G. Drost, S. M. Coelho, Characterization of the
611 transposable element landscape shaping the Ectocarpus genome. *Genome Biol.* 26, 320
612 (2025).

613 47. R. Luthringer, *et al.*, The pseudoautosomal regions of the U/V sex chromosomes of the
614 brown alga ectocarpus exhibit unusual features. *Mol. Biol. Evol.* 32, 2973–2985 (2015).

615 48. P. Liu, *et al.*, 3D chromatin maps of a brown alga reveal U/V sex chromosome spatial
616 organization. *Nature Communications* 15, 9590 (2024).

617 49. S. Liu, *et al.*, Haplotype-resolved genomes provide insights into the origins and functional
618 significance of genome diversity in bivalves. *Cell Rep.* 44, 115697 (2025).

619 50. K. S. Bloom, Centromeric heterochromatin: The primordial segregation machine. *Annu.*
620 *Rev. Genet.* 48, 457–484 (2014).

621 51. J. T. Irelan, E. U. Selker, Cytosine methylation associated with repeat-induced point
622 mutation causes epigenetic gene silencing in Neurospora crassa. *Genetics* 146, 509–523
623 (1997).

624 52. R. R. L. Guillard, P. E. Hargraves, Stichochrysis immobilis is a diatom, not a chrysophyte.
625 *Phycologia* 32, 234–236 (1993).

626 53. S. Chen, Y. Zhou, Y. Chen, J. Gu, Fastp: An ultra-fast all-in-one FASTQ preprocessor.
627 *Bioinformatics* 34, 884–890 (2018).

628 54. S. Andrews, FastQC. *Babraham Bioinformatics* (2010).

629 55. P. Ewels, M. Magnusson, S. Lundin, M. Käller, MultiQC: Summarize analysis results for
630 multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048 (2016).

631 56. E. Kopylova, L. Noé, H. Touzet, SortMeRNA: Fast and accurate filtering of ribosomal
632 RNAs in metatranscriptomic data. *Bioinformatics* 28, 3211–3217 (2012).

633 57. F. O. Glöckner, The SILVA Database Project: An ELIXIR core data resource for high-
634 quality ribosomal RNA sequences. *Biodiversity Information Science and Standards* 3,
635 e36125 (2019).

636 58. A. Dobin, *et al.*, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21
637 (2013).

638 59. C. Bachy, *et al.*, Transcriptional responses of the marine green alga Micromonas pusilla
639 and an infecting prasinovirus under different phosphate conditions. *Environ. Microbiol.* 20,
640 2898–2912 (2018).

641 60. F. X. Ferrer-González, *et al.*, Bacterial transcriptional response to labile exometabolites
642 from photosynthetic picoeukaryote Micromonas commoda. *ISME Communications* 3, 5
643 (2023).

644 61. P. J. Keeling, *et al.*, The Marine Microbial Eukaryote Transcriptome Sequencing Project
645 (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through
646 Transcriptome Sequencing. *PLoS Biol.* 12, e1001889 (2014).

647 62. M. Kolmogorov, J. Yuan, Y. Lin, P. A. Pevzner, Assembly of long, error-prone reads using
648 repeat graphs. *Nat. Biotechnol.* 37, 540–546 (2019).

649 63. C. Camacho, *et al.*, BLAST+: Architecture and applications. *BMC Bioinformatics* 10, 421
650 (2009).

651 64. J. M. Belton, *et al.*, Hi-C: A comprehensive technique to capture the conformation of
652 genomes. *Methods* 58, 268–276 (2012).

653 65. N. Servant, *et al.*, HiC-Pro: An optimized and flexible pipeline for Hi-C data processing.
654 *Genome Biol.* 16, 259 (2015).

655 66. P. Dl Tommaso, *et al.*, Nextflow enables reproducible computational workflows. *Nat.*
656 *Biotechnol.* 35, 316–319 (2017).

657 67. P. A. Ewels, *et al.*, The nf-core framework for community-curated bioinformatics pipelines.
658 *Nat. Biotechnol.* 38, 276–278 (2020).

659 68. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9,
660 357–359 (2012).

661 69. N. Abdennur, L. A. Mirny, Cooler: scalable storage for Hi-C data and other genomically
662 labeled arrays. *Bioinformatics* 36, 311–316 (2020).

663 70. N. Abdennur, *et al.*, Cooltools: Enabling high-resolution Hi-C analysis in Python. *PLoS*
664 *Comput. Biol.* 20, e1012067 (2024).

665 71. S. Wingett, *et al.*, HiCUP: Pipeline for mapping and processing Hi-C data. *F1000Res.* 4,
666 1310 (2015).

667 72. T. Yang, *et al.*, HiCRep: assessing the reproducibility of Hi-C data using a stratum-
668 adjusted correlation coefficient. *Genome Res.* 27, 1939–1949 (2017).

669 73. N. C. Durand, *et al.*, Juicer Provides a One-Click System for Analyzing Loop-Resolution
670 Hi-C Experiments. *Cell Syst.* 3, 95–98 (2016).

671 74. A. Roayaei Ardakany, H. T. Gezer, S. Lonardi, F. Ay, Mustache: Multi-scale detection of
672 chromatin loops from Hi-C and Micro-C maps using scale-space representation. *Genome*
673 *Biol.* 21, 256 (2020).

674 75. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–
675 3100 (2018).

676 76. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–
677 2079 (2009).

678 77. F. Ramírez, F. Dündar, S. Diehl, B. A. Grüning, T. Manke, DeepTools: A flexible platform
679 for exploring deep-sequencing data. *Nucleic Acids Res.* 42, 187–191 (2014).

680 78. D. Buitrago, *et al.*, Nucleosome Dynamics: a new tool for the dynamic analysis of
681 nucleosome positioning. *Nucleic Acids Res.* 47, 9511–9523 (2019).

682 79. F. Krueger, S. R. Andrews, Bismark: A flexible aligner and methylation caller for Bisulfite-
683 Seq applications. *Bioinformatics* 27, 1571–1572 (2011).

684 80. D. T. Schultz, *et al.*, Ancient gene linkages support ctenophores as sister to other animals.
685 *Nature* 618, 110–117 (2023).

686 81. B. Buchfink, K. Reuter, H. G. Drost, Sensitive protein alignments at tree-of-life scale using
687 DIAMOND. *Nat. Methods* 18, 366–368 (2021).

688 82. J. Abramson, *et al.*, Accurate structure prediction of biomolecular interactions with
689 AlphaFold 3. *Nature* 630, 493–500 (2024).

690 83. M. van Kempen, *et al.*, Fast and accurate protein structure search with Foldseek. *Nat.*
691 *Biotechnol.* 42, 243–246 (2024).

692 84. S. Ou, *et al.*, Benchmarking transposable element annotation methods for creation of a
693 streamlined, comprehensive pipeline. *Genome Biol.* 20, 275 (2019).

694 85. A. MacDougall, *et al.*, UniRule: A unified rule resource for automatic annotation in the
695 UniProt knowledgebase. *Bioinformatics* 36, 4643–4648 (2020).

696 86. M. C. Chambers, *et al.*, A cross-platform toolkit for mass spectrometry and proteomics.
697 *Nat. Biotechnol.* 30, 918–920 (2012).

698 87. H. Tsugawa, *et al.*, MS-DIAL: Data-independent MS/MS deconvolution for comprehensive
699 metabolome analysis. *Nat. Methods* 12, 523–526 (2015).

700 88. Z. Pang, *et al.*, MetaboAnalyst 6.0: towards a unified platform for metabolomics data
701 processing, analysis and interpretation. *Nucleic Acids Res.* 52, 398–406 (2024).

702 89. K. Dührkop, *et al.*, SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite
703 structure information. *Nat. Methods* 16, 299–302 (2019).

704
705

Figures and Tables

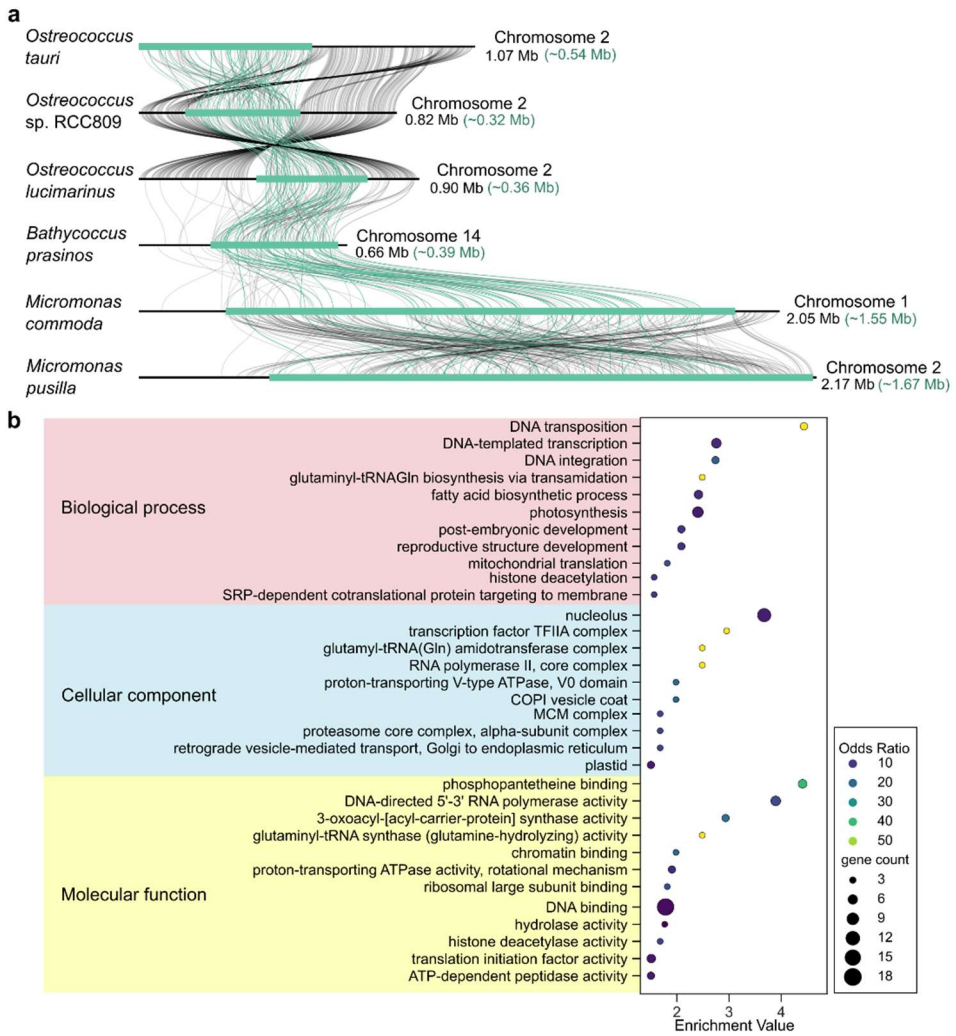


Figure 1. Comparative genomic analysis of BOC1 regions across Mamiellales species. (**a**), Syntenic analysis and gene homology visualisation of BOC1 regions across six Mamiellales species. Connecting lines indicate homologous genes identified through reciprocal best BLASTP hits between *O. tauri* RCC4221, *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 $-\mathrm{Log}_{10}$(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.

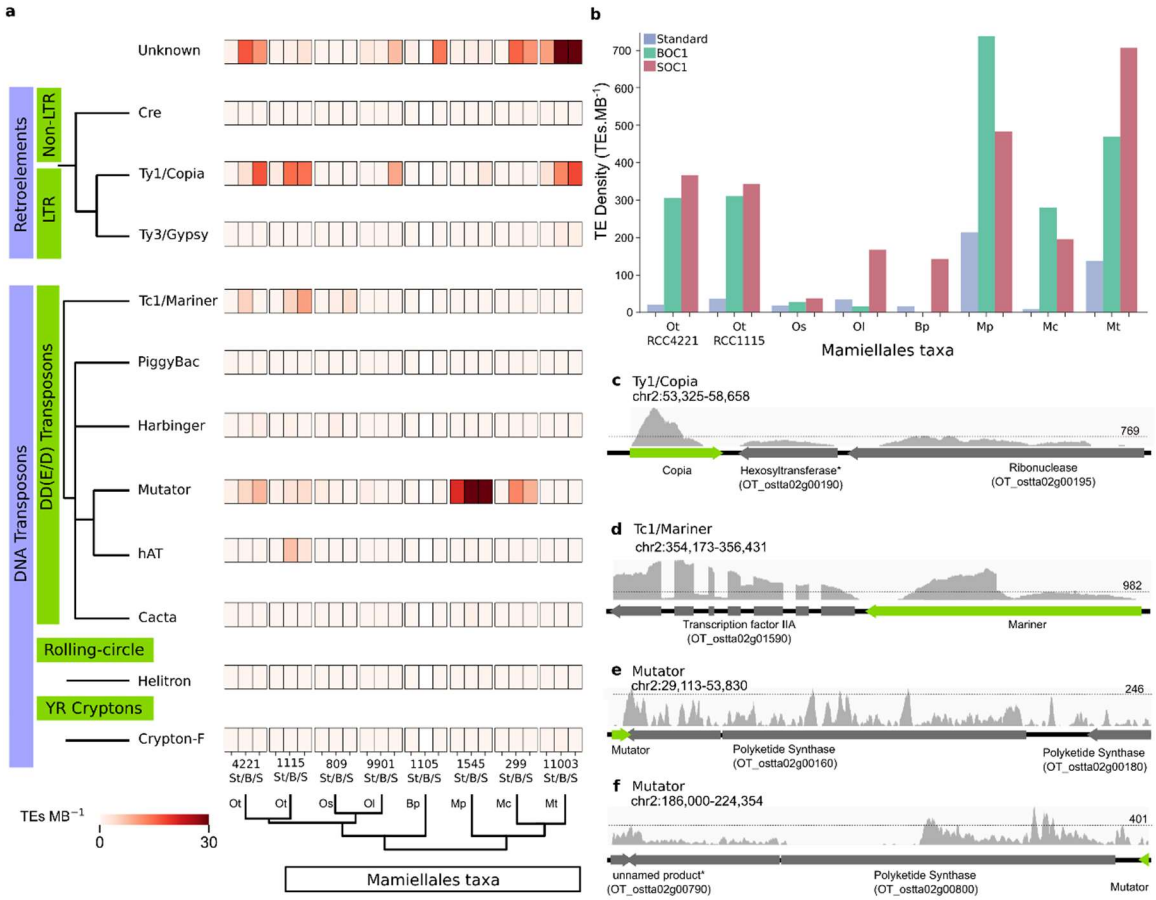
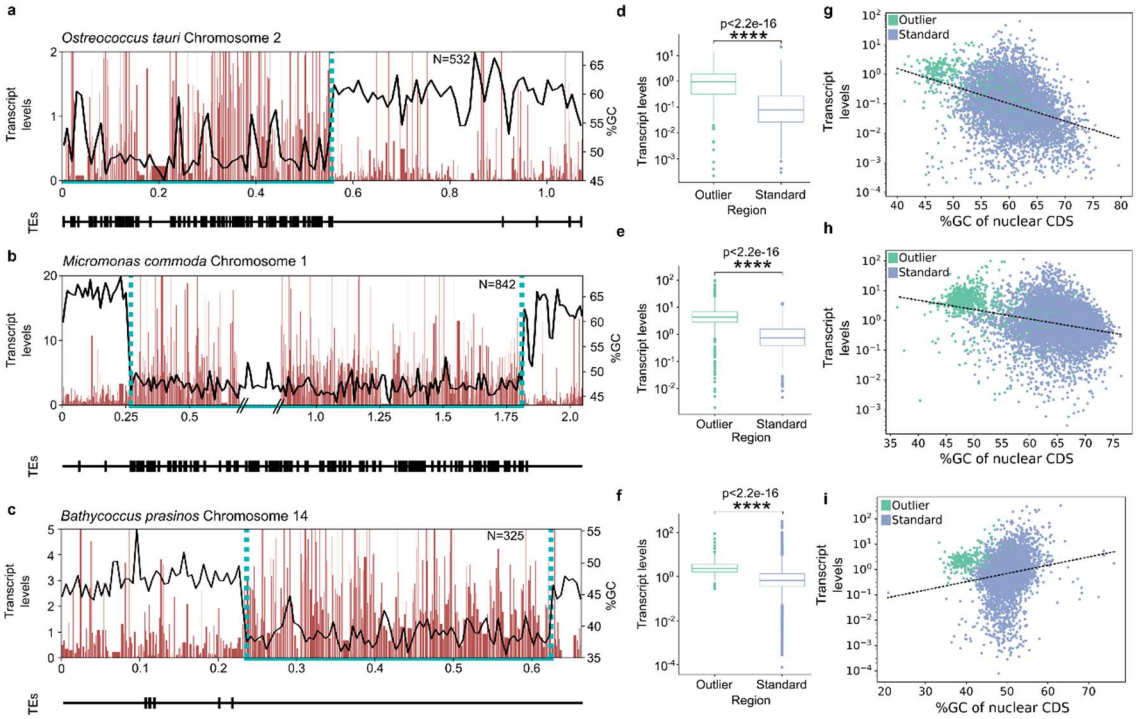


Figure 2. **Distribution and expression of transposable elements across Mamiellales genomes.** (**a**), Dendrogram of major transposable element (TE) classes present in Mamiellales. The left sidebars indicate the classification of elements into retrotransposon, with further subdivision into LTR/non-LTR 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).



741

742 **Figure 3. Higher gene expression in BOC1.** (**a—c**), Expression levels, GC content and TE
743 content across BOC of (**a**) *O. tauri* (RCC4221; chromosome 2), (**b**), *M. commoda* (RCC299;
744 chromosome 1), and (**c**) *B. prasinos* (RCC1105; chromosome 14). Boundary points of the
745 BOC1 regions are labelled with cyan dotted lines; transposable elements are shown on the
746 lower track. (**d—f**), Comparative analysis of gene expression in outlier regions and standard
747 genome; box plots display the distribution of transcript levels across outlier (BOC1 and
748 SOC1) and standard regions for each taxon (Kruskal-Wallis, $p<0.001$). (**g—i**), Correlation
749 between gene expression and GC content of coding sequences (CDS) in (**g**), *O. tauri*
750 (Spearman's rank, $\rho=-0.366$, $p=7.82e^{-242}$), (**h**), *M. commoda* ($\rho = -0.285$, $p = 3.70e^{-192}$), and (**i**)
751 *B. prasinos* ($\rho = -0.38302$, $p = 8.49e^{-13}$); CDS in BOC1 and non-BOC1 regions are represented
752 in green and blue, respectively.
753

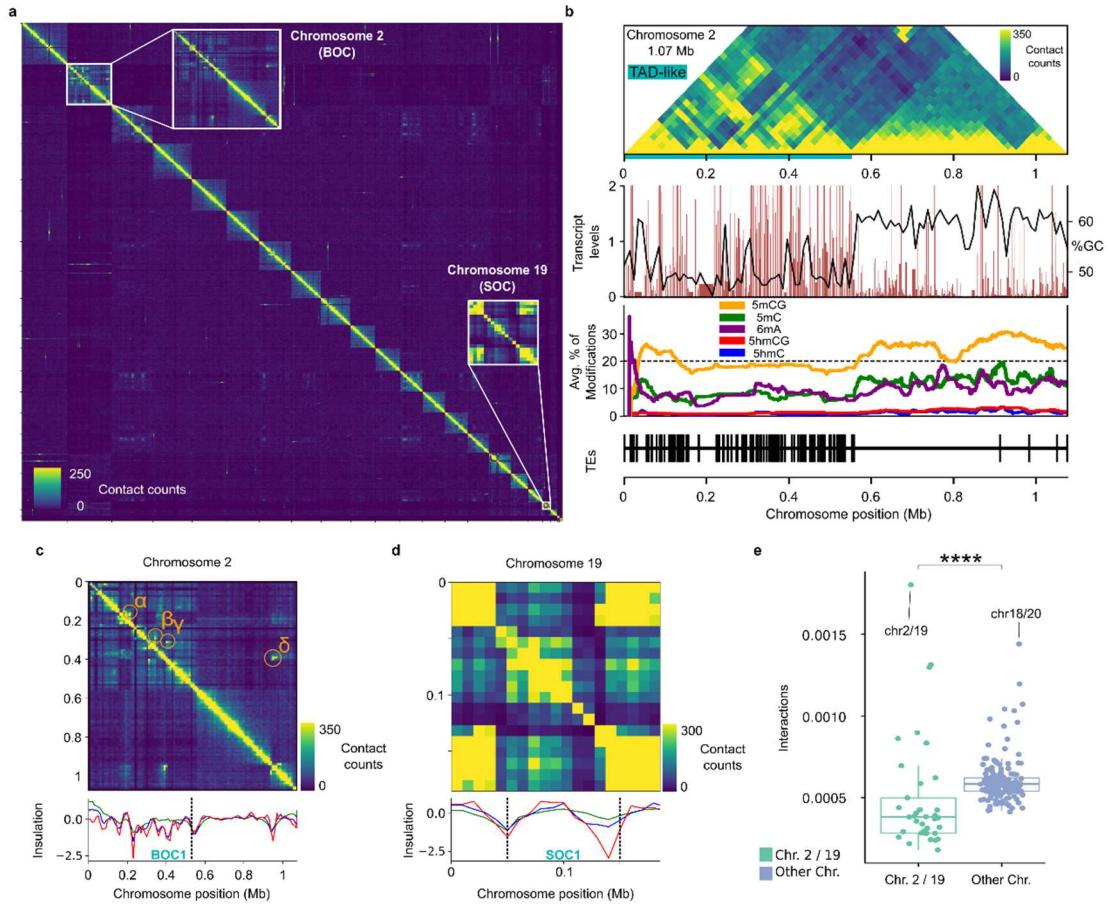


Figure 4. **Chromosomal organisation and three-dimensional genome architecture in *O. tauri*.** (**a**), Hi-C contact map of the *O. tauri* genome binned at 10 kb resolution with normalised interactions counts; chromosomes 2 and 19, which display enriched intrachromosomal contacts, highlighted by white 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.

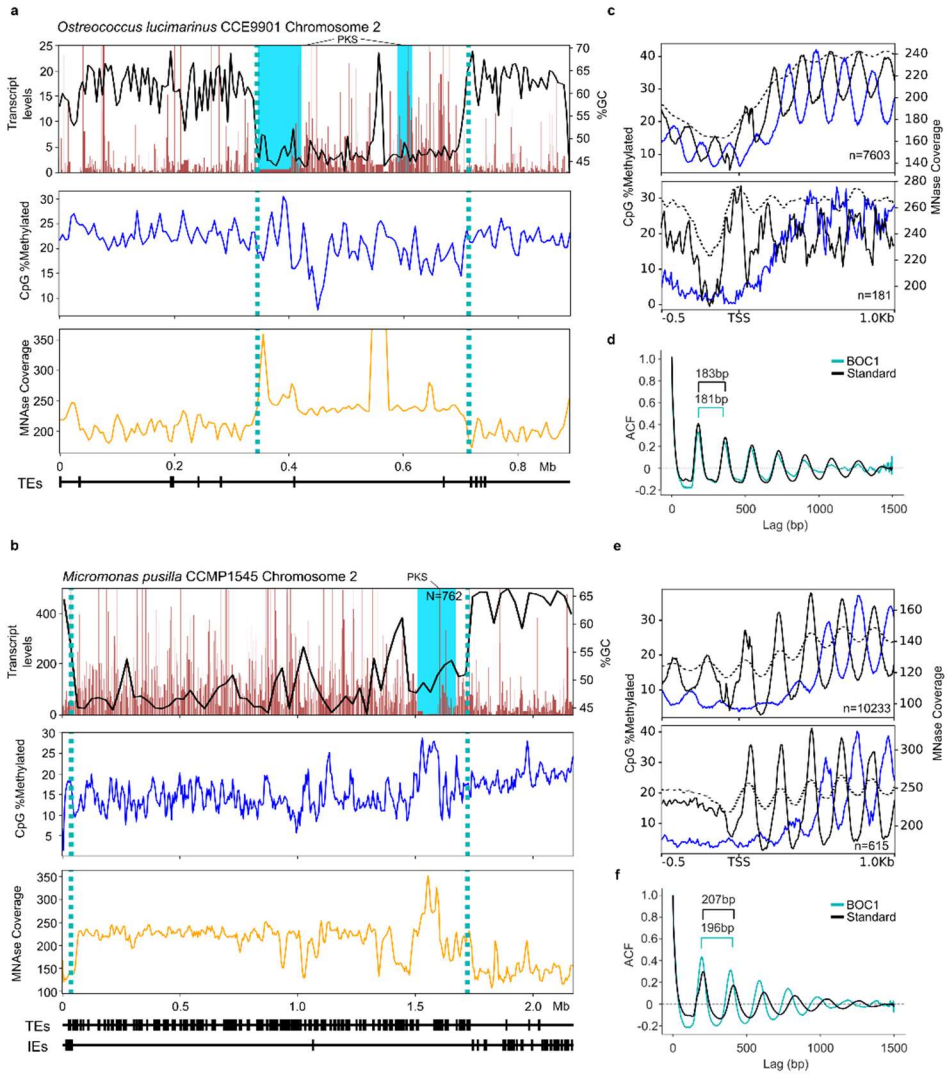


Figure 5. **MNase and CpG profiles across BOC1 in *Ostreococcus lucimarinus* CCE9901 and *Micromonas pusilla* CCMP1545.** (**a, b**), Transcript levels, CpG methylation as a percentage of base coverage which is methylated, MNase coverage alongside TEs and IEs in *O. lucimarinus* CCE9901 (**a**) and *M. 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 centre counts. Light grey is MNase coverage. (**d, f**), The autocorrelation function estimates of nucleosome centre 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.

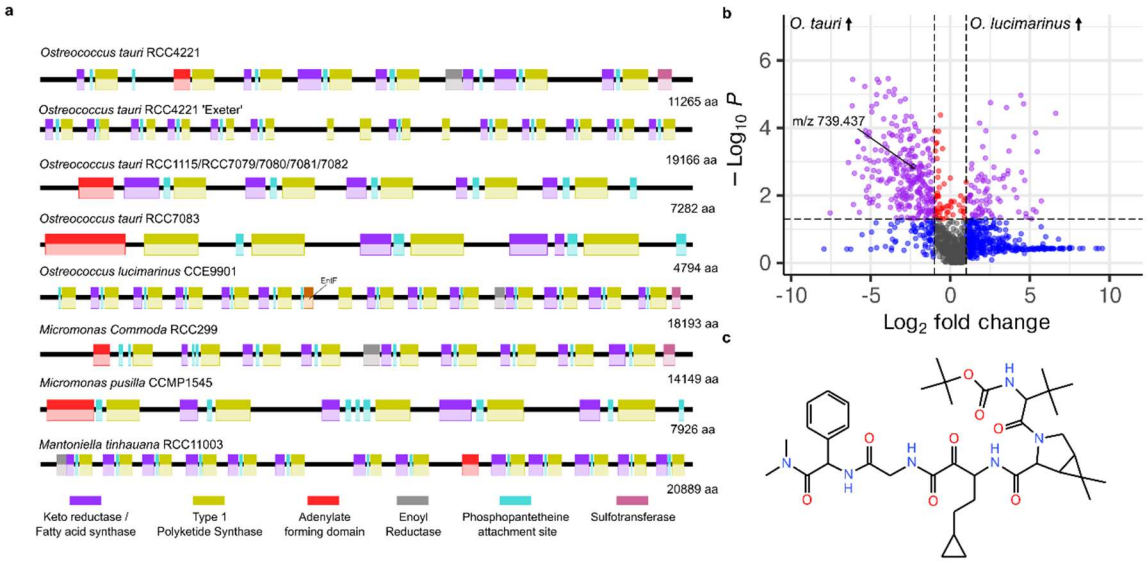


Figure 6. Modular Giant Type I Polyketide synthases show structural diversity across Mamiellales species. (**a**), Schematic representation of the largest orthologous polyketide synthases identified across Mamiellales species, illustrating large size variation in protein product, from 4,794 amino acid 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 1,318 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 Dataset S16–18 for additional structural and metabolomic analyses). (**c**), Structure of the putative polyketide-like compound inferred from LC–MS/MS fragmentation patterns.