'Discovery Report'

A functional bacterial-derived restriction modification system in the mitochondrion of a heterotrophic protist

David S. Milner[a,1], Jeremy G. Wideman[b,c,1,*], Courtney W. Stairs[d], Cory D. Dunn[e] and Thomas A. Richards[a,*]

[a]Department of Zoology, University of Oxford, 11a Mansfield Road, Oxford OX1 3SZ, UK

[b]Biodesign Center for Mechanisms of Evolution, School of Life Sciences, Arizona State University, Tempe, AZ, 85287, USA.

[c]Wissenschaftskolleg zu Berlin, Wallotstraße 19, 14193, Berlin, Germany.

[d]Department of Biology, Lund University, Sölvegatan 35, 223 62 Lund, Sweden

[e]Institute of Biotechnology, Helsinki Institute of Life Science, University of Helsinki, Helsinki, 00014, Finland

[1]These authors contributed equally to this work.

[*]Corresponding authors:

Thomas A. Richards, Department of Zoology, University of Oxford, 11a Mansfield Road, Oxford OX1 3SZ, United Kingdom. E-mail: thomas.richards@zoo.ox.ac.uk

Jeremy G. Wideman, Biodesign Center for Mechanisms of Evolution, School of Life Sciences, Arizona State University, Tempe, AZ, 85287, USA. Email: jeremy.wideman@asu.edu

Short title: A functional mitochondrial-encoded restriction modification system in a heterotrophic protist.

Abstract

The overarching trend in mitochondrial evolution is functional streamlining coupled with gene loss; therefore, gene acquisition by mitochondria is considered to be exceedingly rare. Selfish elements in the form of self-splicing introns occur in many organellar genomes, but the wider diversity of selfish elements, and how they persist in organellar genomes, has not been explored. In the mitochondrial genome of a marine heterotrophic katablepharid protist, we identify a functional type II restriction modification system originating from a horizontal gene transfer event involving bacteria related to flavobacteria. This restriction modification system consists of an HpaII-like endonuclease and a cognate cytosine methyltransferase. We demonstrate that these proteins are functional by heterologous expression in both bacterial and eukaryotic cells. These results suggest that toxin-antitoxin selfish elements, such as restriction modification systems, could be co-opted by eukaryotic genomes to drive uniparental organellar inheritance.

Introduction

Endosymbiosis, the localization and functional integration of one cell within another [1–3], can lead to the evolution of specialized organellar compartments responsible for a range of cellular and biochemical functions [4]. Mitochondria and plastids originate from endosymbiotic events, and typically retain vestigial genomes of bacterial ancestry [5,6]. While sequencing initiatives have demonstrated that mitochondrial gene content can vary extensively, their evolution in every eukaryotic lineage is typified by both functional and genomic reduction [7,8]. Rare gene replacements and novel gene acquisitions into mitochondrial genomes have been identified, particularly involving plant-to-plant gene transfers [9–12], with plants also susceptible to the transfer of entire organellar genomes [13–15]. In addition, mitochondrial group I and II self-splicing introns demonstrate a pattern of mosaic distribution consistent with multiple recent gene transfer and loss events [16,17]. Chloroplasts generally exhibit the same reductive evolutionary trends [18]; and although horizontal gene transfer (HGT) of a bacterial operon into the chloroplast genome of eustigmatophyte algae (Ochrophyta), including *Monodopsis* and *Vischeria,* has been reported, the functional relevance of this acquisition is not yet clear [19]. While novel functional genes have entered chloroplast genomes, and replacement genes can find their way into mitochondrial DNA (mtDNA), no gain-of-function transfers into the mitochondrial genome have, to our knowledge, been reported previously.

Truly ‘selfish’ genetic elements serve no function except to replicate themselves [20], even at the cost of host fitness. However, some horizontally transferred selfish elements have been co-opted to perform critical functions in host cells [21]. For example, type II restriction modification (RM) systems can provide host cells with protection from invasion by viruses, plasmids, or other sources of foreign DNA [22]. RM selfish elements work by the coordinated regulation of two enzymes that behave as a type IV toxin-antitoxin system [23]; the restriction endonuclease acts as a ‘toxin’ by cutting DNA at specific recognition sequence motifs, while a methyltransferase acts as an ‘anti-toxin’ by modifying nucleotides at the same recognition sequence, thereby protecting the DNA from cleavage by the endonuclease. If found within an organellar genome, a functional RM system could simply act to protect the genome from invasion by viruses/phages, plasmids, or other sources of foreign DNA. However, RM systems

in organelles could also act as a strong 'gene drive', ensuring that a single mitochondrial haplotype would quickly sweep to fixation in a sexual population via mitochondrial fusion events.

We have recently explored the content of diverse protist mitochondrial genomes using targeted culture-independent single cell approaches [24]. This process allowed us to recover the first complete mitochondrial genomes from katablepharid protists. Here, we describe the identification and characterization of four open reading frames (ORFs) comprising two type II RM selfish elements within katablepharid mtDNA. To our knowledge, we present the first example of a type II RM selfish element within any eukaryotic genome (nuclear or organellar), that likely derives from an HGT event into the mitochondrial genome. We report the phylogenetic ancestry of these genes and assess the activity of the encoded enzymes by heterologous expression in *Escherichia coli* and *Saccharomyces cerevisiae*. We suggest that these mitochondrial-encoded proteins may constitute a hitherto undescribed system controlling patterns of organelle inheritance.

Results

Identification of unique restriction-modification selfish elements in katablepharid mitochondrial genomes

Our recent initiative to assess mitochondrial genome content using environmentally sampled protistan single-cell amplified genome (SAG) sequencing resulted in the complete mitochondrial sequence of multiple marine heterotrophic katablepharid protists [24]. The contemporary publication of the complete *Leucocryptos marina* mitochondrial genome confirmed the identity of the single-cell amplified genome (SAG)-derived mitochondrial DNAs as katablepharids [25]. The genomes from the SAG-generated mtDNA and *Leucocryptos* mtDNA were identical in their repertoires of canonical mitochondrial genes including tRNA genes (**Fig. 1a**). They shared synteny throughout the majority of the genome, including near identical intron locations in *rnl*, *cob*, and *cox1* (grey in **Fig. 1a**), and retained three unassigned ORFs at identical genomic locations (orange in **Fig. 1a**). The regions lacking synteny between the complete genomes encode *atp9*, *rns*, the vast majority of *tRNA*s, as well as a variety of unassigned ORFs. Of these eight unassigned ORFs in this region of the SAG-derived katablepharid mitochondrial genome, three had homologues in *L. marina* (orange in **Fig. 1a**),

but five did not retrieve *L. marina* proteins as top hits using BLAST searches (red in **Fig. 1a**). One of these was a highly divergent GIY-YIG homing endonuclease, two were identified as restriction enzymes, and two were identified as cytosine methyltransferases (CMs) (**Fig. 1b**). The restriction enzymes and CMs comprised two tandemly-encoded type II Restriction-Modification (RM) selfish genetic elements each consisting of a restriction endonuclease and a cognate CM. Specifically, the two katablepharid RMs were identified to be composed of a HpaII restriction endonuclease (Kat-HpaII) and its cognate cytosine methyltransferase (Kat-HpaII-CM), and a MutH/Restriction endonuclease type II (Kat-MutH) with its cognate cytosine-C5 methyltransferase (Kat-MutH-CM) (**Fig. 1b**). The Kat-HpaII RM system is flanked by near-identical (152/155 bp) sequences that may reflect the recent integration of this selfish element (shown in orange in **Fig. 1b**) into the katablepharid mtDNA.

BLASTP analysis of the putative methyltransferases against the REBASE database (http://rebase.neb.com; accessed January 2021) indicated that Kat-HpaII-CM is likely specific to a CCGG DNA recognition sequence, with an *Algibacter* methyltransferase (accession: ALJ03853.1) as the top hit (59% identity). Furthermore, BLASTP analysis suggested that the Kat-MutH-CM was likely specific to a GATC DNA recognition sequence, and the top hit was a methyltransferase from *Arenitalea lutea* (genome accession: ALIH01000012.1, 74% identity). Analysis of the putative endonucleases also suggested that Kat-HpaII was specific for a CCGG DNA recognition sequence, with an *Aggregatibacter* endonuclease (accession: RDE88890.1) as the top BLASTP hit (36% identity), and that the Kat-MutH endonuclease may be specific for GATC, recovering as the top BLASTP hit a DNA mismatch repair protein from *Mangrovimonas* (accession: KFB02001.1, 27% identity). The 2240 bp region encoding the Kat-HpaII/Kat-HpaII-CM RM system lacks any CCGG motifs, which are expected to occur by chance once every 752 bp in the katablepharid mitochondrion, based on a GC content of 38% for this organelle genome. Taken together, our analysis predicts that the gene products within each katablepharid RM pair target the same recognition sequences.

To confirm that all identified regions were not the result of contamination or genome assembly artifacts, we re-amplified the corresponding region of the mitochondrion from the SAG DNA samples, confirming the four-gene architecture of the selfish element identified was present in two samples ('katablepharid 1 (i.e. 'K1')' and 'katablepharid 4' (i.e. 'K4')) (**Fig. 1b**),

and that the genes are adjacent to the katablepharid mitochondrial *atp9* and *rns* genes. This PCR analysis also confirmed the existence of a reduced variant (in 'katablepharid 3'), which consists of only the N-terminus region of the MutH-cytosine methyltransferase gene (**Fig. 1b**).

Next, we conducted three separate, targeted PCRs using environmental DNA samples recovered from parallel marine water samples collected on the same date, and from the same site, as those which contained the individual cells sorted for genome sequencing [24]. These analyses further confirmed that the selfish elements were found adjacent to mitochondrial genes, and that the products were not an artefact of multiple displacement amplification as part of the single-cell sequencing pipeline. We also identified an additional contig possessing an intermediate reduced form of the selfish element gene architecture, which contained only the MutH-CM gene (**Fig. 1b**). In total, the integration of RM systems into mitochondrial genomes was independently confirmed five times. Collectively, these results indicate that RM selfish genetic elements have been incorporated into mtDNAs, and have been subjected to rapid evolutionary change, including gene loss/ORF-reduction.

To explore if the identified selfish element genes are expressed, we interrogated a collection of marine meta-transcriptome data publicly available at the Ocean Gene Atlas (OGA, available at http://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/). We identified a number of eukaryotic transcripts from geographically diverse marine sampling sites with strong nucleotide identity to the *kat-HpaII*, *kat-HpaII-CM*, and *kat-MutH-CM* genes (**Supplementary File S1**), demonstrating that this selfish genetic element is expressed from the katablepharid mitochondrial genome. Regions identified in the MATOU_v1_metaT transcriptome database with >95% identity are shown in **Fig. 1b**. Interestingly, two of the OGA RNAseq derived contigs that showed >99% nucleotide identity to the Kat-HpaII mitochondrial gene were composed of sequence reads sampled from multiple sites in the Pacific, Southern Atlantic, and Indian Oceans, and the Mediterranean Sea. These samples included both 'surface' and 'deep chlorophyll maximum' samples. These results suggest that the Kat-HpaII endonuclease is transcriptionally active across a wide range of ocean environments. Furthermore, the OGA transcript sequences included one contig that traverses the *trnW* gene and the repetitive flanking sequence upstream of Kat-HpaII-CM/Kat-MutH-CM gene cluster, indicative of mitochondrial co-transcription.

Katablepharid mitochondrial RM system has flavobacterial ancestry

To explore the phylogenetic ancestry of the selfish genetic element we conducted phylogenetic analysis using Bayesian and maximum likelihood approaches, with a focus upon the complete Kat-HpaII-CM and Kat-HpaII found in the K4 assembly [24], as Kat-MutH-CM and Kat-MutH were found to have no detectable function (discussed below). The phylogeny of the restriction endonuclease showed limited bootstrap support, with the mitochondrial genes branching with weak bootstrap support within the flavobacteria (**Fig. S1a**). In contrast, phylogenetic analysis of Kat-HpaII-CM (**Fig. S1b**) and the concatenated alignment of both Kat-HpaII and Kat-HpaII-CM demonstrated strong bootstrap support (**Fig. 2a**) for the mitochondrial selfish genetic element branching within a clade of flavobacteria. There is currently no evidence that flavobacteria, or genetic material derived from the flavobacteria, played a role in the origin of the eukaryotes or the mitochondrial organelle [26,27]. Furthermore, the katablepharid SAG assemblies contained no obvious contaminating flavobacterial-like sequences (**Fig. S2**). As such, we conclude that the selfish genetic element is a recent transfer to the mitochondrial genome from a donor species that branches within, or close to, the flavobacteria. Our phylogenetic analysis demonstrated that the selfish genetic element represented a relatively extended branch in the phylogeny (**Fig. 2a**), suggesting evolutionary scenarios consistent with invasion of the mtDNA genome, such as population bottlenecking, positive selection, or relaxed selection.

To further explore the nature of sequence evolution associated with this HGT event, we calculated the codon usage frequencies of the Kat4-HpaII RM, the conserved protein-coding gene repertoire of the Kat4 mtDNA, and the *Algibacter*-HpaII RM. Using Fisher Exact tests, we demonstrated that codon usage was significantly different for 14 amino acids when comparing the Kat4 complement of unambiguously ancestral Kat4 mitochondrial proteins and the *Algibacter*-HpaII RM. In contrast, the Kat4-HpaII RM represents an intermediate, with 6 amino acids with codon usage differing from the Kat4 mtDNA [again sampling all unambiguously ancestral Kat4 mitochondrial proteins], and 7 amino acids with codon usage differing from that of *Algibacter*-HpaII RM (**Fig. 2b**; see **Supplementary File S3** for raw data). These data are consistent with the hypothesis that the Kat4-HpaII RM is in the process of domestication towards the sequence characteristics of the host mtDNA. Such changes may

also be, in part, a driver and/or consequence of the accelerated evolutionary rate indicated by the relatively long branch the katablepharid selfish element forms in the phylogenetic trees (**Fig. 2a**)

Confirmation of a functional mitochondrial katablepharid methyltransferase

To explore the function of the selfish element and to test if it has undergone pseudogenization, we cloned the Kat-HpaII-CM and its closest bacterial homologue in terms of sequence identity, *Algibacter* HpaII-CM (Alg-HpaII-CM), into plasmid pACYC184 and expressed them in *E. coli* Top10 cells. This *E. coli* strain does not contain any methyltransferases that target the putative HpaII-CM recognition sequence (CCGG), but instead expresses Dcm methylase, which methylates the second cytosine residue in CCWGG [28], and Dam methylase, which methylates adenine residues in the sequence GATC [29]. The HpaII-CM-expressing *E. coli* Top10 strains were cultured for 16 h alongside a control strain harbouring an empty plasmid. These plasmids were then extracted, linearized and subject to bisulfite conversion, a process that converts cytosine nucleotides to uracil but does not alter methylated 5-methylcytosines (5-mC). A 299 bp region of each plasmid was PCR amplified and sequenced. In the plasmid from the control strain, all CCGG sites appeared as TTGG in sequencing chromatograms, whereas plasmids sequenced from strains expressing Kat-HpaII-CM or Alg-HpaII-CM contained TCGG sites (**Fig. 3a**), demonstrating the ability of Kat-HpaII-CM to methylate CCGG sequences at the second cytosine base.

To confirm Kat-HpaII-CM function, we transformed the plasmid expressing this katablepharid sequence into *E. coli* strain DH5α, harbouring the C^{me}CGG-cutting enzyme McrA [30], and into the *mcrA*- *E. coli* Top10 strain. Comparisons of transformation efficiency confirmed that the katablepharid methyltransferase is toxic in an *E. coli* DH5α background (**Fig. 3b**), further demonstrating that Kat-HpaII-CM encodes a functional enzyme which methylates CCGG sites.

Next, we also performed bisulfite conversion and sequencing experiments using Kat-MutH-CM, targeting an alternative 233 bp region of the plasmid to enable detection of potential methylation at GATC sites. However, we found no evidence of any methylation at GATC, or other sites, by Kat-MutH-CM, suggesting that this enzyme may have lost its GATC

specific catalytic activity, requires additional factors for its function, or has gained an alternative function.

Confirmation of a functional mitochondrial katablepharid endonuclease

To explore the function of the candidate endonucleases, we cloned the putative *Algibacter* HpaII endonuclease (Alg-HpaII), the katablepharid MutH-like endonuclease (Kat-MutH), and Kat-HpaII into a pBAD expression vector, transformed these plasmids into *E. coli*, and compared culture growth for each resulting strain. The strain expressing Kat-MutH showed no evidence of toxicity, demonstrating a similar growth dynamic to the vector-only *E. coli* strain (**Fig. S3**), and the functions of Kat-MutH-CM and Kat-MutH were not pursued further. In contrast, cultures of strains expressing the Kat-HpaII and Alg-HpaII grew slowly, consistent with these genes encoding functional endonucleases that constitute a *bona fide* 'toxin' (**Fig. 3c**). The katablepharid HpaII showed a greater potency during these experiments when compared to the *Algibacter* HpaII. In order to explore if the Kat-HpaII and Kat-HpaII-CM function as a toxin/anti-toxin pair, we co-expressed these two proteins. This demonstrated that Kat-HpaII-CM was able to partially reverse the effects of Kat-HpaII expression in *E. coli* (**Fig. 3d**). Subsequent experiments increasing the expression of the Kat-HpaII enzyme by removal of an additional ATG at the 5' of the sequence led to this rescue being perturbed (**Fig. S4**), suggesting that differences in the relative expression of the toxin/antitoxin can determine the degree of toxicity.

Targeting of HpaII-CM and HpaII to yeast mitochondria confirms methyltransferase and endonuclease activities

To further explore the likely roles of Kat-HpaII-CM and Kat-HpaII in katablepharid mitochondria, we targeted each protein to the mitochondria of *S. cerevisiae* cells using an amino-terminal Su9 mitochondrial targeting sequence (MTS) from *Neurospora crassa* [31]. Constructs also contained a carboxyl-terminal GFP tag to allow confirmation of mitochondrial localization, and proteins were controlled by a galactose-inducible promoter to allow temporal induction of gene expression. Following induction of su9(MTS)-Kat-HpaII-CM-GFP, we sequenced a region of the mitochondrial *COX1* gene following bisulfite conversion. As seen following heterologous expression in *E. coli*, CCGG sites of mtDNA were methylated, indicating that Kat-HpaII-CM could function in the context of a mitochondrial matrix (**Fig. 4a**).

We also sequenced this region of the *COX1* gene from a *S. cerevisiae* isolate lacking the su9(MTS)-Kat-HpaII-CM-GFP plasmid and demonstrated that these residues are not methylated in wild-type cultures. We assessed whether su9(MTS)-Kat-HpaII-CM-GFP would be recruited to mtDNA by staining mitochondrial nucleoids with DAPI [32]. This demonstrated that the katablepharid methyltransferase co-localised with punctate DAPI foci (**Fig. 4b**).

When su9(MTS)-Kat-HpaII-GFP was targeted to mitochondria, this protein was also found in puncta, yet the DAPI appeared absent, indicating that the mtDNA has likely been degraded (**Fig. 4b**). This ability of Kat-HpaII to damage mtDNA was suggested by an increase in the formation of petite colonies after su9(MTS)-Kat-HpaII-GFP induction (**Fig. 4c**). Taken together, our results indicate that the Kat-HpaII-CM and Kat-HpaII are able to function within mitochondria to methylate and degrade mitochondrial DNA.

Discussion

Here, we have revealed the integration of a functional type II RM system into the mtDNA of a katablepharid – the first known instance detected within a eukaryotic genome. We confirm that Kat-HpaII and Kat-HpaII-CM are functional when expressed in both prokaryotic and eukaryotic cells, and we provide data demonstrating a toxin/antitoxin functional relationship between these two proteins. Why would this active RM selfish element reside within a mitochondrial genome? We suggest several possible evolutionary scenarios. First, the katablepharid type II RM selfish element could simply represent a recent invasion of no functional or evolutionary consequence for its katablepharid host. We do see evidence of RM system degeneration within some of the katablepharid mitochondrial genome sequences sampled, suggesting that selection for maintenance of this selfish element is patchy, and loss is tolerated.

Second, the Kat-HpaII and Kat-HpaII-CM system may protect the mitochondria from foreign DNA. In bacteria, RM selfish elements are thought to function as a defence against foreign unmethylated DNA [33], such as viruses/phages and plasmids, which are also known to invade mitochondria [34,35]. However, any fitness benefit to cells harbouring these elements related to this function would be conditional upon regular exposure to foreign sources of DNA. Consistent with this proposition, we detected evidence of expression of the

278 Kat-HpaII gene from multiple oceanic environments, implicating a wide biogeographic
279 distribution of active gene transcription.

280 Third, and most intriguing among these possibilities, this RM element may be driving
281 spread of the host mtDNA within the katablepharid population. Previous studies indicate that
282 in a sexual population, selfish mitochondrial mutants spread rapidly, whereas asexual
283 populations are relatively protected from similar patterns of invasions [36–39]. Furthermore,
284 mitochondrial reticulation and/or fusion is documented in many eukaryotes [40,41]. Thus,
285 crosses of Kat4-HpaII RM+ and RM- individuals would hypothetically initially result in a mixed
286 population of mitochondria. After mitochondrial fusion, RM+ mtDNA would lead to digestion
287 of unprotected RM- mtDNA, leading to selfish element-mediated, uniparental inheritance of
288 RM+ mtDNA and, potentially, the rapid spread of this mitochondrial haplotype. To further
289 explore this possibility we searched for genes which putatively encode meiosis components
290 in our four katablepharid SAGs and identified gene fragments of six meiosis-associated
291 proteins (MSH5, XRCC3, DMC1, SPO11, Brambleberry and SNF2; see **Supplementary File S2**)
292 in K2/K4, suggesting that katablepharids, like most eukaryotes, contain meiosis-specific genes
293 (e.g. SPO11; MSH5 [42]), and may be capable of sexual reproduction, although sex has not
294 been directly observed in this lineage [43]. However, these SAGs are incomplete and
295 extremely fragmented [24], and therefore require confirmation with additional data.

296 Uniparental inheritance of cytoplasmic organelles is a consistent trend across diverse
297 eukaryotic groups and has multiple, independent origins [44]. Therefore, the invasion of
298 organellar genomes by RM selfish elements may constitute a hitherto unrecognised
299 mechanism for gene drive that enables differential parental inheritance of mitochondrial
300 genomes, independent of direct nuclear control. While methylation/nuclease functions may
301 contribute to the uniparental inheritance of chloroplasts in *Chlamydomonas* [45,46], the
302 mechanisms in this system are unclear, and the genes responsible have not been reported to
303 be a consequence of an HGT invasion event. Furthermore, the invasion of selfish genetic
304 elements, based on toxin-antitoxin function, into organellar genomes has been predicted
305 [47], although, until now, not identified. This prediction sets out that selfish genetic elements
306 will take up important roles in inter-organellar genome conflict (i.e. a form of organellar
307 'warfare') [47]. It is therefore possible, even likely, that the RM system identified here may

act as weapon in such warfare, manipulating the inheritance patterns of mitochondrial genomes in katablepharids.

Materials and Methods

Phylogenetic analysis of restriction-modification selfish elements encoded in katablepharid mitochondrial genomes

To determine the origins of the katablepharid mitochondrial-encoded RM selfish element we collected putative homologues from the NCBI *nr* database using katablepharid HpaII (Kat-HpaII) and HpaII-CM (Kat-HpaII-CM) as queries. The top hits were predominantly from the Flavobacteriaceae, suggesting that the katablepharid RM originated within this group. To confirm the phylogenetic origins of the katablepharid RM system, we collected protein sequences from diverse bacterial phyla that encoded HpaII and HpaII-CM in tandem, then reconstructed single-gene and concatenated phylogenies. HpaII and HpaII-CM orthologues were aligned with MUSCLE [48] and manually trimmed using Mesquite v.2.75 [49]. The two-gene concatenation was performed by hand in Mesquite v.2.75. Phylogenetic tree reconstructions were performed using MrBayes v.3.2.6 for Bayesian analysis [50] using the following parameters: prset aamodelpr = fixed (WAG); mcmcngen = 2,000,000; samplefreq = 1000; nchains = 4; startingtree = random; sumt burnin = 250. Splits frequencies were checked to ensure convergence. Maximum-likelihood bootstrap values (100 pseudoreplicates) were obtained using RAxML v.8.2.10 [51] under the LG model [52].

Analysis of codon usage

Codon usage frequencies of the proteins encoded by the Kat4 and Algibacter HpaII and HpaII-CM selfish elements, as well as the unambiguously ancestral Kat4 mitochondrial proteins [26], were determined using the Sequence Manipulation Suite server [53]. Amino acid codon usage frequencies were compared using a Fisher Exact test in R (version 1.3.1073) [54].

Identification of katablepharid-related restriction-modification systems in metagenomic databases

All four genes of the two selfish elements were BLASTN searched against the Ocean Gene Atlas [55] (searched December 2020, tool available at: http://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/) OM-RGC_v2_metaT (prokaryote) and

MATOU_v1_metaT (eukaryote) transcriptome databases. Only hits of over 100 bp in length with DNA identity scores in excess of 95% were retained for further analysis (see **Supplementary File S1**).

Identification of putative meiosis protein encoding genes in katablepharid SAGs

Hidden Markov models (HMMs) corresponding to meiosis-associated proteins [56,57] were retrieved from Pfam, PNTHR, EGGNOG and TIGR databases via InterPro (https://www.ebi.ac.uk/interpro/; November, 2020); see **Supplementary File S2** for accession numbers. These HMMs were used as queries against a six-frame translation of the Katablepharid SAGs (K1: sample 11B_35C, K2: 11H_35C, K3: 5F_35A, K4: 6E_35B; https://figshare.com/articles/dataset/Single_Cell_Genomic_Assemblies/7352966) using hmmsearch with an e-value (-E) cut-off of 0.1, with all other parameters at default. The nucleotide sequences from resulting hits were used as queries against the non-redundant (*nr*) database (November, 2020) using BLASTX [58] to allow for intron read-through. If the majority of the top hits against the *nr* database corresponded to the same meiosis-associated protein, then the sequence was included in **Supplementary File S2**.

PCR confirmation of katablepharid restriction-modification selfish elements

To validate the presence of the RM system on the katablepharid mitochondrial genome assembly, and to further assess the katablepharid mitochondrial RM diversity, we conducted PCR, using a range of templates: i) a katablepharid single-cell amplified genome (SAG) DNA from Wideman *et al* [24] and ii) DNA extracted from a water sample taken at a depth of 20 m from the same site, the Monterey Bay Aquarium Research Institute time-series station M2, and on the same date, as the single-cell isolations [24]. PCR amplifications were performed using Phusion polymerase (New England Biolabs) and the primers detailed in **Table S1**. Each 25 µl reaction contained 200 nM of each primer, 400 nM dNTPs and 1 ng template DNA. Cycling conditions were 2 mins at 98°C followed by 30 cycles of 10 s at 98°C, 20 s at 64.3°C, 2-3 mins at 72°C, and a final extension of 7 mins at 72°C. PCR products were purified (GeneJet PCR Purification Kit, Thermo Fisher Scientific), adenosine-tailed using GoTaq Flexi DNA polymerase (Promega), and cloned into pSC-A-amp/kn using a StrataClone PCR Cloning Kit (Agilent Technologies). Plasmids were then Sanger sequenced using T7/T3 primers or the

366 original PCR primers (MWG Eurofins), with additional internal sequencing reactions
367 performed when necessary.

368 **Plasmid construction**

369 Sequences were codon optimised for *E. coli* or *S. cerevisiae* expression and synthesised *de*
370 *novo* (Synbio Tech, NJ). For *E. coli* expression, putative methyltransferases were cloned into
371 the BamHI/SalI sites of the low copy vector pACYC184 (New England Biolabs) with an
372 upstream Shine-Dalgarno consensus sequence (5'-AGGAGG-3'), and putative endonucleases
373 were cloned into the PstI/HindIII sites of pBAD HisA (Thermo Fisher Scientific). For expression
374 of proteins in *S. cerevisiae*, each ORF was fused to an N-terminal Su9 pre-sequence from
375 *Neurospora crassa* for targeting to the mitochondrion, and to a C-terminal GFP tag for
376 visualization by fluorescent microscopy. Kat-HpaII-CM and Kat-HpaII were cloned into the
377 BamHI/KpnI sites of pYX223-mtGFP and pYES-mtGFP plasmids, respectively [31]. All plasmid
378 constructs are detailed in **Table S2**.

379 ***E. coli* transformation and proliferation assays of strains expressing components of the RM**
380 **system**

381 Plasmids containing putative methyltransferase and endonuclease genes were transformed
382 into chemically competent *E. coli* Top10 (*dcm*+ *dam*+, *mcrA*-) or DH5α (*dcm*+ *dam*+, *mcrA*+).
383 Where transformations into DH5α were unsuccessful, biological triplicate transformations
384 were performed into both Top10 and DH5α to assess strain-specific incompatibility. This was
385 achieved by performing transformations where equal concentrations (50 ng) of pDM040 or
386 pACYC184 (empty vector control) were added to each competent cell aliquot, before plating
387 onto LB Cm_{45}, incubating at 37°C for 16 h, then counting colony forming units.

388 To assess proliferation of each *E. coli* Top10 strain, duplicate cultures were grown for
389 16 h at 37°C (200 rpm shaking) in LB Amp_{50} Cm_{45} before being diluted to OD_{600} 0.1 in the same
390 medium. 100 µL of each culture was inoculated into a 96-well plate and incubated at 37°C
391 with 200 rpm double-orbital shaking in a BMG FLUOstar Omega Lite instrument. Proliferation
392 was assessed by measuring OD_{595} at 5-minute intervals for 480 minutes.

393 **Bisulfite conversion to assess for methylase activity**

To assay for 5-methylcytosine (5-mC) methyltransferase activity, *E. coli* Top10 strains with a pACYC184 vector containing Kat-HpaII-CM (pDM040), Kat-MutH-CM (pDM042), or *Algibacter* methyltransferase (Alg-HpaII-CM) (pDM041) were grown for 16 h at 37°C (200 rpm shaking) in LB Cm_{45}. Plasmids were extracted using a GeneJet Plasmid Miniprep kit (Thermo Fisher Scientific), linearised using HindIII to avoid supercoiling, then gel extracted (Promega Wizard SV Gel and PCR Clean-Up System). Linear plasmids were subjected to bisulfite conversion using the EpiMark Bisulfite Conversion Kit (New England Biolabs), following the manufacturer's instructions. A 299 bp region of each plasmid was amplified with primers pACYC184_5mC_F and pACYC184_5mC_R2 to assess CCGG methylation, and a 233 bp region was amplified with primers pACYC184_region2_5mC_F2/R2 to assess GATC methylation. Both primer pairs (**Table S1**) were designed to amplify bisulfite-converted DNA. 25 µl reactions containing 1x GoTaq G2 Hot Start Green Master Mix (Promega), 1 µM each primer and 1 µL of 100-fold diluted plasmid template were used, with the following cycling conditions: 2 mins at 94°C, followed by 35 cycles of 15 s at 94°C, 30 s at 50°C and 30 s at 72°C, then a final extension of 5 mins at 72°C. PCR products were then purified (Promega Wizard SV Gel and PCR Clean-Up System) and sequenced on both strands (Eurofins Genomics) to identify bases which remained as cytosines, indicative of a 5-mC modification at this site.

To assess mtDNA methylation in *S. cerevisiae* cells, 1 mL of culture was purified using a Promega Wizard genomic DNA purification kit, following the manufacturer's instructions for isolating genomic DNA from yeast. Bisulfite conversion was performed as above, with 500 ng of genomic DNA used in each reaction. Primers cox1_bisulfite_F and cox1_bisulfite_R (**Table S1**) were then used to amplify a 443 bp region of the mitochondrial *cox1* gene using GoTaq Hot Start Master Mix (Promega). Each 50 µL reaction contained 500 nM each primer and cycling conditions were as described above. PCR products were purified using a Wizard PCR clean-up kit (Promega) before sequencing (Eurofins Genomics).

GFP localisation of heterologously expressed RM system components in yeast using spinning disc confocal microscopy

Plasmids pDM071, encoding su9(MTS)-Kat-HpaII-GFP, or pDM072, encoding su9(MTS)-Kat-HpaII-CM-GFP (**Table S2**), were transformed into competent *S. cerevisiae* BY4742 cells, using the method described by Thompson *et al* [59], and selected on Scm-ura [0.69% yeast nitrogen

base without amino acids (Formedium), 770 mg L^{-1} complete supplement mix (CSM) lacking uracil (Formedium), 2% (wt/vol) glucose, and 1.8% (wt/vol) Agar No. 2 Bacteriological (Neogen)] or Scm-his agar (containing CSM-histidine in place of CSM-uracil), respectively. Cells were grown for 16 h in Scm-his or Scm-ura plus 2% glucose at 30°C, diluted 10-fold, and induced in fresh media containing 2% galactose (instead of glucose) for 12 h. Cells were then harvested by centrifugation, and resuspended in sterile water containing 1 µg mL^{-1} DAPI for 30 min, which preferentially labels mtDNA in the absence of fixation [32]. Cells were then observed by spinning-disc confocal microscopy using an Olympus IX81 inverted microscope affixed with a CSU-X1 Spinning Disk unit (Yokogawa) and 405 nm/488 nm lasers.

Assessment of petite formation in *S. cerevisiae*

A *S. cerevisiae* W303 derivative, CAY169 [60] harbouring plasmid pDM071 (HpaII) was grown to mid-logarithmic phase in Scm-ura [0.69% yeast nitrogen base without amino acids (Formedium), 770 mg L^{-1} complete supplement mix lacking uracil (Formedium), 2% (wt/vol) glucose, 20 mg L^{-1} adenine sulfate]. Cells were pelleted and re-suspended in fresh media containing 2% galactose (instead of glucose) for 12 h to induce expression of su9(MTS)-Kat-HpaII-GFP, then plated on Scm-ura agar (as above, containing glucose as the sole carbon source). Plates were incubated for 3 days at 30°C and imaged to assess the formation of petite colonies following the pulse of mitochondrial Kat-HpaII expression.

Data Release Statement

The single-cell amplified genome assemblies and the K1, K3 and K4 mitochondrial genome contigs, originally from Wideman ***et al.***, 2020, are available at https://doi.org/10.6084/m9.figshare.7352966 and https://doi.org/10.6084/m9.figshare.7314728, respectively. All additional relevant data are within the paper and its Supporting Information files.

Acknowledgements

The authors would like to thank Dayana Salas-Leiva for helpful advice about meiosis and Alexandra Worden's Group for assistance in provision of samples for the initial study (Wideman *et al.* 2020). pYX223-mtGFP and pYES-mtGFP were a gift from Benedikt Westermann (Addgene plasmid # 45051/45053). The Wissenschaftskolleg zu Berlin provided

accommodation for a joint lab meeting at which this project was devised. EMBO provided travel arrangements enabling the joint lab meeting to occur. C.D.D. is supported by the Sigrid Juselius Foundation, the Academy of Finland (331556), and the Jane and Aatos Erkko Foundation. T.A.R. is supported by a Royal Society University Research Fellowship (UF130382) and additional awards through the EMBO YIP program. C.W.S is supported by a Vetenskaprådet starting grant (2020-05071).

References

1. Bonen L, Doolittle WF. On the prokaryotic nature of red algal chloroplasts. Proceedings of the National Academy of Sciences. 1975;72: 2310–2314. doi:10.1073/pnas.72.6.2310

2. Gray MW, Doolittle WF. Has the endosymbiont hypothesis been proven? Microbiol Rev. 1982;46: 1–42.

3. Gray MW. The Bacterial Ancestry of Plastids and Mitochondria. BioScience. 1983;33: 693–699. doi:10.2307/1309349

4. Gabaldón T, Huynen MA. From Endosymbiont to Host-Controlled Organelle: The Hijacking of Mitochondrial Protein Synthesis and Metabolism. PLOS Computational Biology. 2007;3: e219. doi:10.1371/journal.pcbi.0030219

5. Burger G, Gray MW, Lang BF. Mitochondrial genomes: anything goes. Trends in Genetics. 2003;19: 709–716. doi:10.1016/j.tig.2003.10.012

6. Green BR. Chloroplast genomes of photosynthetic eukaryotes. The Plant Journal. 2011;66: 34–44. doi:https://doi.org/10.1111/j.1365-313X.2011.04541.x

7. Adams KL, Palmer JD. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Molecular Phylogenetics and Evolution. 2003;29: 380–395. doi:10.1016/S1055-7903(03)00194-5

8. Johnston IG, Williams BP. Evolutionary Inference across Eukaryotes Identifies Specific Pressures Favoring Mitochondrial Gene Retention. Cell Syst. 2016;2: 101–111. doi:10.1016/j.cels.2016.01.013

9. Rice DW, Alverson AJ, Richardson AO, Young GJ, Sanchez-Puerta M V., Munzinger J, et al. Horizontal Transfer of Entire Genomes via Mitochondrial Fusion in the Angiosperm Amborella. Science. 2013;342: 1468–1473. doi:10.1126/science.1246275

10. Goremykin V V., Salamini F, Velasco R, Viola R. Mitochondrial DNA of Vitis vinifera and the issue of rampant horizontal gene transfer. Molecular Biology and Evolution. 2009. doi:10.1093/molbev/msn226

11. Mower JP, Stefanović S, Hao W, Gummow JS, Jain K, Ahmed D, et al. Horizontal acquisition of multiple mitochondrial genes from a parasitic plant followed by gene conversion with host mitochondrial genes. BMC Biology. 2010. doi:10.1186/1741-7007-8-150

12. Hao W, Richardson AO, Zheng Y, Palmer JD. Gorgeous mosaic of mitochondrial genes created by horizontal transfer and gene conversion. PNAS. 2010;107: 21576–21581. doi:10.1073/pnas.1016295107

13. Stegemann S, Bock R. Exchange of Genetic Material Between Cells in Plant Tissue Grafts. Science. 2009;324: 649–651. doi:10.1126/science.1170397

14. Gurdon C, Svab Z, Feng Y, Kumar D, Maliga P. Cell-to-cell movement of mitochondria in plants. Proc Natl Acad Sci USA. 2016;113: 3395–3400. doi:10.1073/pnas.1518644113

15. Hertle AP, Haberl B, Bock R. Horizontal genome transfer by cell-to-cell travel of whole organelles. Sci Adv. 2021;7: eabd8215. doi:10.1126/sciadv.abd8215

16. Hausner G. Introns, Mobile Elements, and Plasmids. Organelle Genetics: Evolution of Organelle Genomes and Gene Expression. 2012. pp. 329–357. doi:10.1007/978-3-642-22380-8_13

17. Wu B, Hao W. Horizontal Transfer and Gene Conversion as an Important Driving Force in Shaping the Landscape of Mitochondrial Introns. G3 (Bethesda). 2014;4: 605–612. doi:10.1534/g3.113.009910

18. Martin W, Herrmann RG. Gene Transfer from Organelles to the Nucleus: How Much, What Happens, and Why? Plant Physiology. 1998;118: 9–17. doi:10.1104/pp.118.1.9

19. Yurchenko T, Ševčíková T, Strnad H, Butenko A, Eliáš M. The plastid genome of some eustigmatophyte algae harbours a bacteria-derived six-gene cluster for biosynthesis of a novel secondary metabolite. Open Biol. 2016;6. doi:10.1098/rsob.160249

20. Bravo Núñez MA, Nuckolls NL, Zanders SE. Genetic Villains: Killer Meiotic Drivers. Trends Genet. 2018;34: 424–433. doi:10.1016/j.tig.2018.02.003

21. Werren JH. Selfish genetic elements, genetic conflict, and evolutionary innovation. PNAS. 2011;108: 10863–10870. doi:10.1073/pnas.1102343108

22. Vasu K, Nagaraja V. Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. Microbiol Mol Biol Rev. 2013;77: 53–72. doi:10.1128/MMBR.00044-12

23. Mruk I, Kobayashi I. To be or not to be: regulation of restriction–modification systems and other toxin–antitoxin systems. Nucleic Acids Res. 2014;42: 70–86. doi:10.1093/nar/gkt711

24. Wideman JG, Monier A, Rodríguez-Martínez R, Leonard G, Cook E, Poirier C, et al. Unexpected mitochondrial genome diversity revealed by targeted single-cell genomics

519 of heterotrophic flagellated protists. Nature Microbiology. 2020;5: 154–165.
520 doi:10.1038/s41564-019-0605-4

521 25. Nishimura Y, Kume K, Sonehara K, Tanifuji G, Shiratori T, Ishida K, et al. Mitochondrial
522 Genomes of Hemiarma marina and Leucocryptos marina Revised the Evolution of
523 Cytochrome c Maturation in Cryptista. Front Ecol Evol. 2020;8.
524 doi:10.3389/fevo.2020.00140

525 26. Roger AJ, Muñoz-Gómez SA, Kamikawa R. The Origin and Diversification of Mitochondria.
526 Current Biology. 2017;27: R1177–R1192. doi:10.1016/j.cub.2017.09.015

527 27. Fan L, Wu D, Goremykin V, Xiao J, Xu Y, Garg S, et al. Phylogenetic analyses with systematic
528 taxon sampling show that mitochondria branch within Alphaproteobacteria. Nature
529 Ecology & Evolution. 2020;4: 1213–1219. doi:10.1038/s41559-020-1239-x

530 28. Buryanov YI, Bogdarina IG, Bayev AA. Site specificity and chromatographic properties of
531 E. coli K12 and EcoRII DNA-cytosine methylases. FEBS LETTERS. 1978;88: 4.

532 29. Geier GE, Modrich P. Recognition sequence of the dam methylase of Escherichia coli K12
533 and mode of cleavage of Dpn I endonuclease. J Biol Chem. 1979;254: 1408–1413.

534 30. Raleigh EA, Trimarchi R, Revel H. Genetic and Physical Mapping of the Mcra (Rgla) and
535 Mcrb (Rglb) Loci of Escherichia Coli K-12. Genetics. 1989;122: 279–296.

536 31. Westermann B, Neupert W. Mitochondria-targeted green fluorescent proteins:
537 convenient tools for the study of organelle biogenesis in Saccharomyces cerevisiae.
538 Yeast. 2000;16: 1421–1427. doi:https://doi.org/10.1002/1097-
539 0061(200011)16:15<1421::AID-YEA624>3.0.CO;2-U

540 32. Williamson DH, Fennell DJ. [62] Visualization of yeast mitochondrial DNA with the
541 fluorescent stain “DAPI.” Methods in Enzymology. Academic Press; 1979. pp. 728–733.
542 doi:10.1016/0076-6879(79)56065-0

543 33. Kobayashi I. Behavior of restriction-modification systems as selfish mobile elements and
544 their impact on genome evolution. Nucleic acids research. 2001;29: 3742–56.

545 34. Hillman BI, Cai G. The Family Narnaviridae: Simplest of RNA Viruses. In: Ghabrial SA,
546 editor. Advances in Virus Research. Academic Press; 2013. pp. 149–176.
547 doi:10.1016/B978-0-12-394315-6.00006-4

548 35. Nishimura Y, Shiratori T, Ishida K, Hashimoto T, Ohkuma M, Inagaki Y. Horizontally-
549 acquired genetic elements in the mitochondrial genome of a centrohelid Marophrys sp.
550 SRT127. Scientific Reports. 2019;9: 4850. doi:10.1038/s41598-019-41238-6

551 36. Futcher B, Reid E, Hickey DA. Maintenance of the 2 micron circle plasmid of
552 Saccharomyces cerevisiae by sexual transmission: an example of a selfish DNA. Genetics.
553 1988;118: 411–415.

554 37. Zeyl C, Bell G, Green DM. Sex and the Spread of Retrotransposon Ty3 in Experimental
555 Populations of Saccharomyces Cerevisiae. Genetics. 1996;143: 1567–1577.

556 38. Goddard MR, Godfray HCJ, Burt A. Sex increases the efficacy of natural selection in
557 experimental yeast populations. Nature. 2005;434: 636–640. doi:10.1038/nature03405

558 39. Harrison E, MacLean RC, Koufopanou V, Burt A. Sex drives intracellular conflict in yeast.
559 Journal of Evolutionary Biology. 2014;27: 1757–1763.
560 doi:https://doi.org/10.1111/jeb.12408

561 40. Kawano S, Takano H, Kuroiwa T. Sexuality of Mitochondria: Fusion, Recombination, and
562 Plasmids. In: Jeon KW, Jarvik J, editors. International Review of Cytology. Academic
563 Press; 1995. pp. 49–110. doi:10.1016/S0074-7696(08)62496-1

564 41. Logan DC. The dynamic plant chondriome. Seminars in Cell & Developmental Biology.
565 2010;21: 550–557. doi:10.1016/j.semcdb.2009.12.010

566 42. Schurko AM, Logsdon JM. Using a meiosis detection toolkit to investigate ancient asexual
567 "scandals" and the evolution of sex. BioEssays. 2008;30: 579–589.
568 doi:https://doi.org/10.1002/bies.20764

569 43. Speijer D, Lukeš J, Eliáš M. Sex is a ubiquitous, ancient, and inherent attribute of
570 eukaryotic life. Proc Natl Acad Sci U S A. 2015;112: 8827–8834.
571 doi:10.1073/pnas.1501725112

572 44. Birky CW. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms
573 and evolution. Proceedings of the National Academy of Sciences. 1995;92: 11331–
574 11338. doi:10.1073/pnas.92.25.11331

575 45. Sager R, Lane D. Molecular Basis of Maternal Inheritance. Proceedings of the National
576 Academy of Sciences. 1972;69: 2410–2413. doi:10.1073/pnas.69.9.2410

577 46. Nishimura Y. An mt+ gamete-specific nuclease that targets mt- chloroplasts during sexual
578 reproduction in C. reinhardtii. Genes & Development. 2002;16: 1116–1128.
579 doi:10.1101/gad.979902

580 47. Havird JC, Forsythe ES, Williams AM, Werren JH, Dowling DK, Sloan DB. Selfish
581 Mitonuclear Conflict. Current Biology. 2019;29: R496–R511.
582 doi:10.1016/j.cub.2019.03.020

583 48. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput.
584 Nucleic Acids Research. 2004;32: 1792–1797. doi:10.1093/nar/gkh340

585 49. Maddison WP, Maddison DR. Mesquite 2.75: a modular system for evolutionary analysis.
586 2011.

587 50. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2:
588 Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space.
589 Systematic Biology. 2012;61: 539–542. doi:10.1093/sysbio/sys029

590 51. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
591 phylogenies. Bioinformatics. 2014;30: 1312–1313. doi:10.1093/bioinformatics/btu033

592 52. Le SQ, Gascuel O. An Improved General Amino Acid Replacement Matrix. Molecular
593 Biology and Evolution. 2008;25: 1307–1320. doi:10.1093/molbev/msn067

594 53. Stothard P. The Sequence Manipulation Suite: JavaScript Programs for Analyzing and
595 Formatting Protein and DNA Sequences. BioTechniques. 2000;28: 1102–1104.
596 doi:10.2144/00286ir01

597 54. R Core Team. R: A language and environment for statistical computing. R Foundation for
598 Statistical Computing, Vienna, Austria. 2013. Available: http://www.R-project.org/.

599 55. Villar E, Vannier T, Vernette C, Lescot M, Cuenca M, Alexandre A, et al. The Ocean Gene
600 Atlas: exploring the biogeography of plankton genes online. Nucleic Acids Research.
601 2018;46: W289–W295. doi:10.1093/nar/gky376

602 56. Malik S-B, Pightling AW, Stefaniak LM, Schurko AM, Jr JML. An Expanded Inventory of
603 Conserved Meiotic Genes Provides Evidence for Sex in Trichomonas vaginalis. PLOS ONE.
604 2008;3: e2879. doi:10.1371/journal.pone.0002879

605 57. Hofstatter PG, Lahr DJG. All Eukaryotes Are Sexual, unless Proven Otherwise. BioEssays.
606 2019;41: 1800246. doi:https://doi.org/10.1002/bies.201800246

607 58. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool.
608 Journal of Molecular Biology. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2

609 59. Thompson JR, Register E, Curotto J, Kurtz M, Kelly R. An improved protocol for the
610 preparation of yeast cells for transformation by electroporation. Yeast. 1998;14: 565–
611 571. doi:https://doi.org/10.1002/(SICI)1097-0061(19980430)14:6<565::AID-
612 YEA251>3.0.CO;2-B

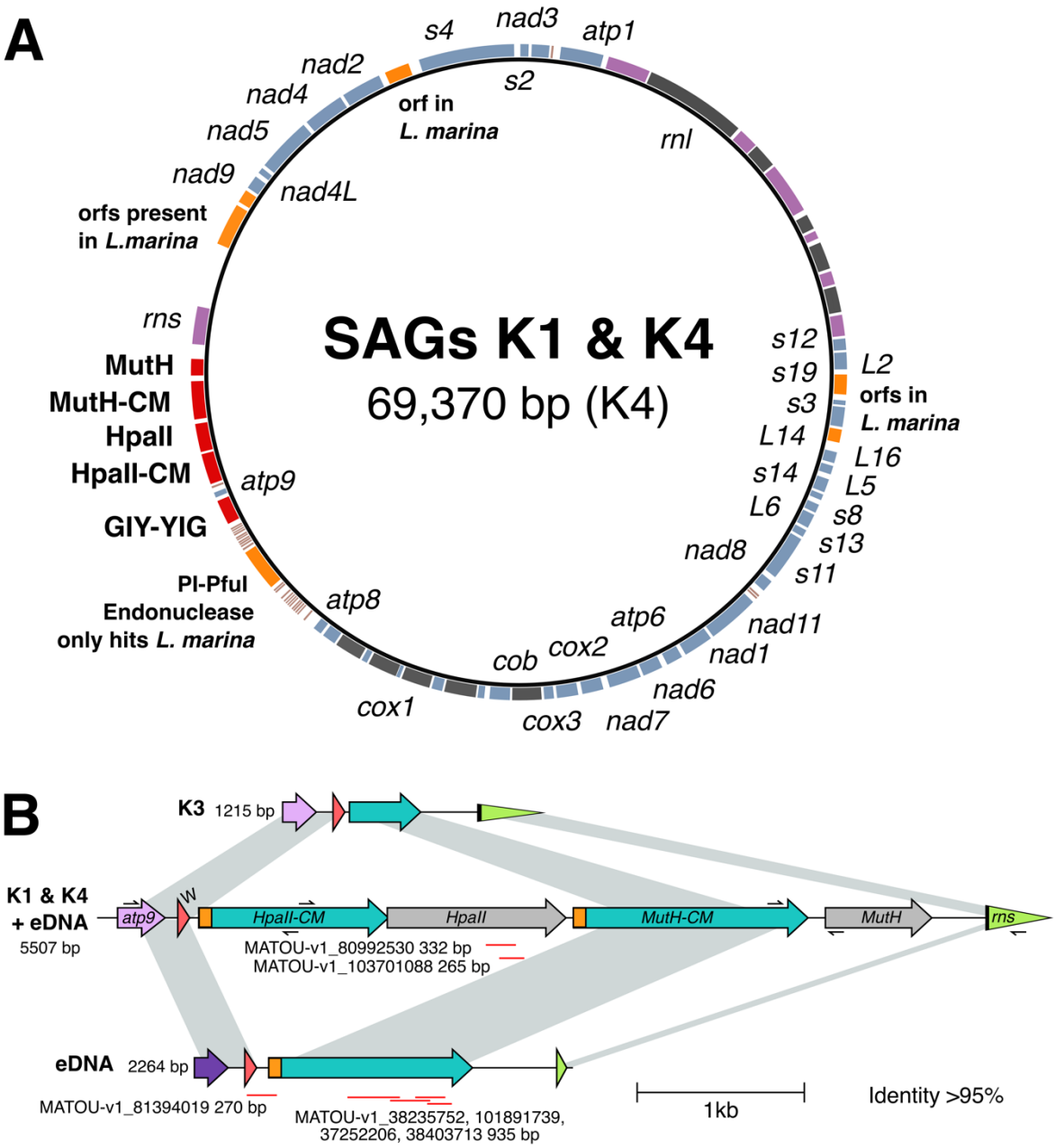
613 60. Milner DS, Attah V, Cook E, Maguire F, Savory FR, Morrison M, et al. Environment-
614 dependent fitness gains can be driven by horizontal gene transfer of transporter-
615 encoding genes. PNAS. 2019;116: 5613–5622. doi:10.1073/pnas.1815994116

616 61. Gilchrist CLM, Chooi Y-H. clinker & clustermap.js: Automatic generation of gene cluster
617 comparison figures. bioRxiv. 2020 [cited 19 Nov 2020]. doi:10.1101/2020.11.08.370650

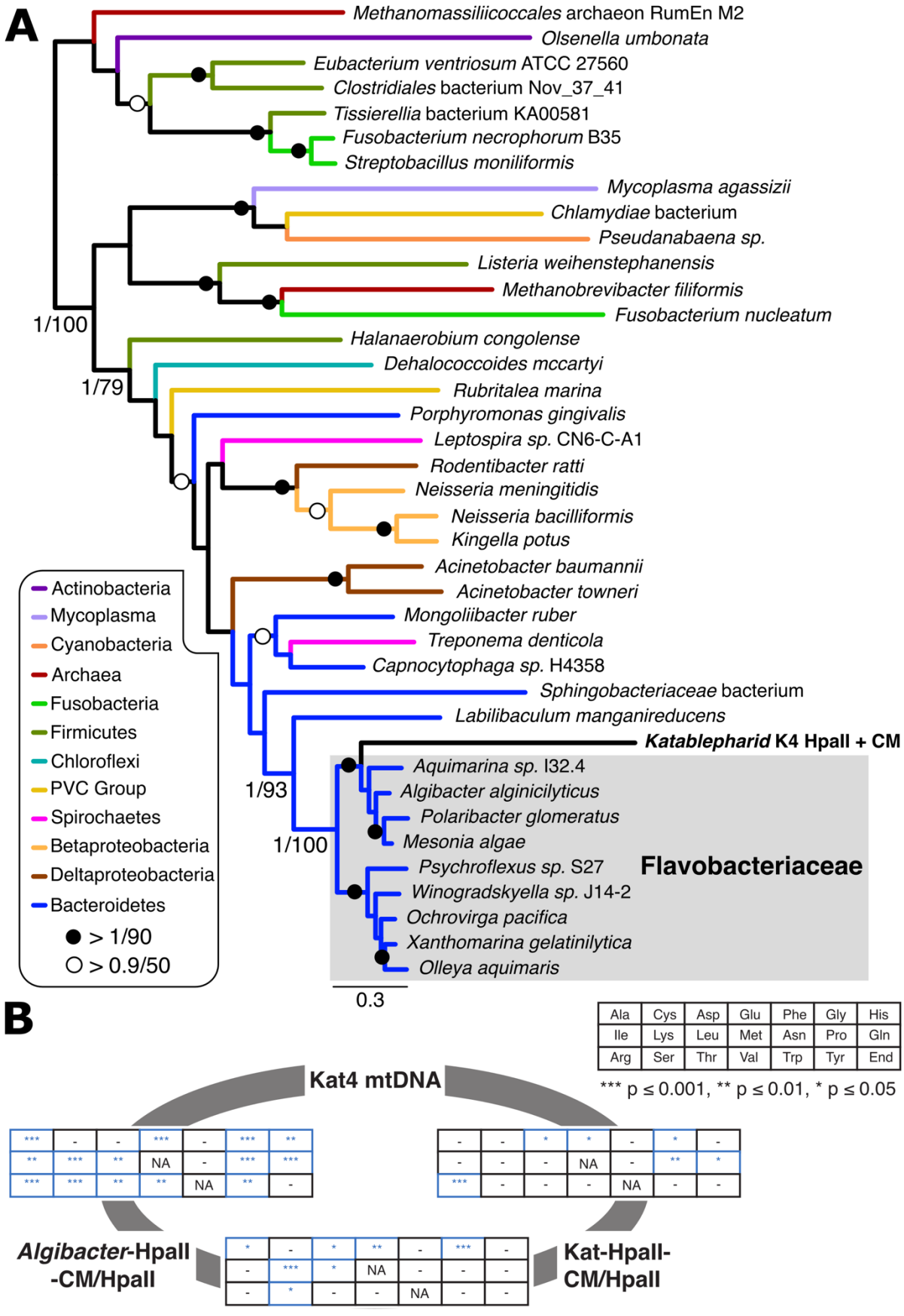
618 62. Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M. Blobology: exploring raw genome
619 data for contaminants, symbionts and parasites using taxon-annotated GC-coverage
620 plots. Frontiers in Genetics. 2013;4: 237. doi:10.3389/fgene.2013.00237

621

Figure 1

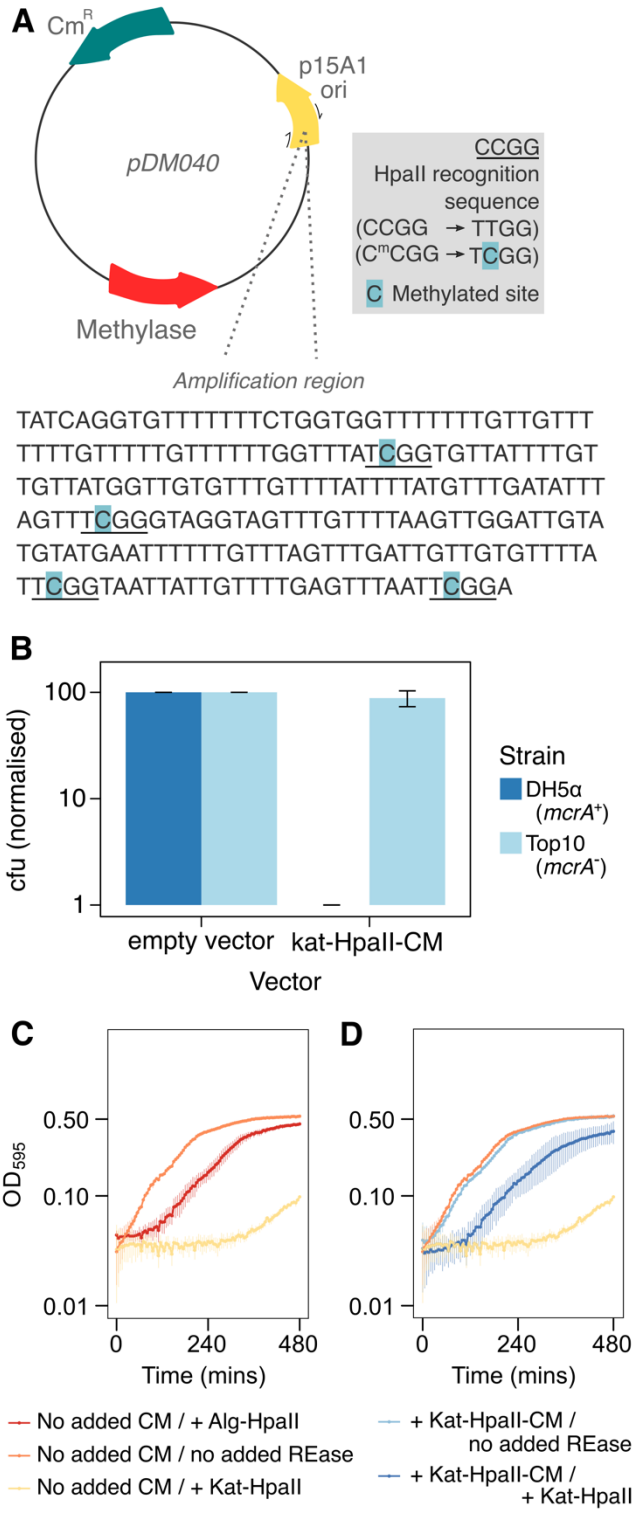


627 **Figure 2**

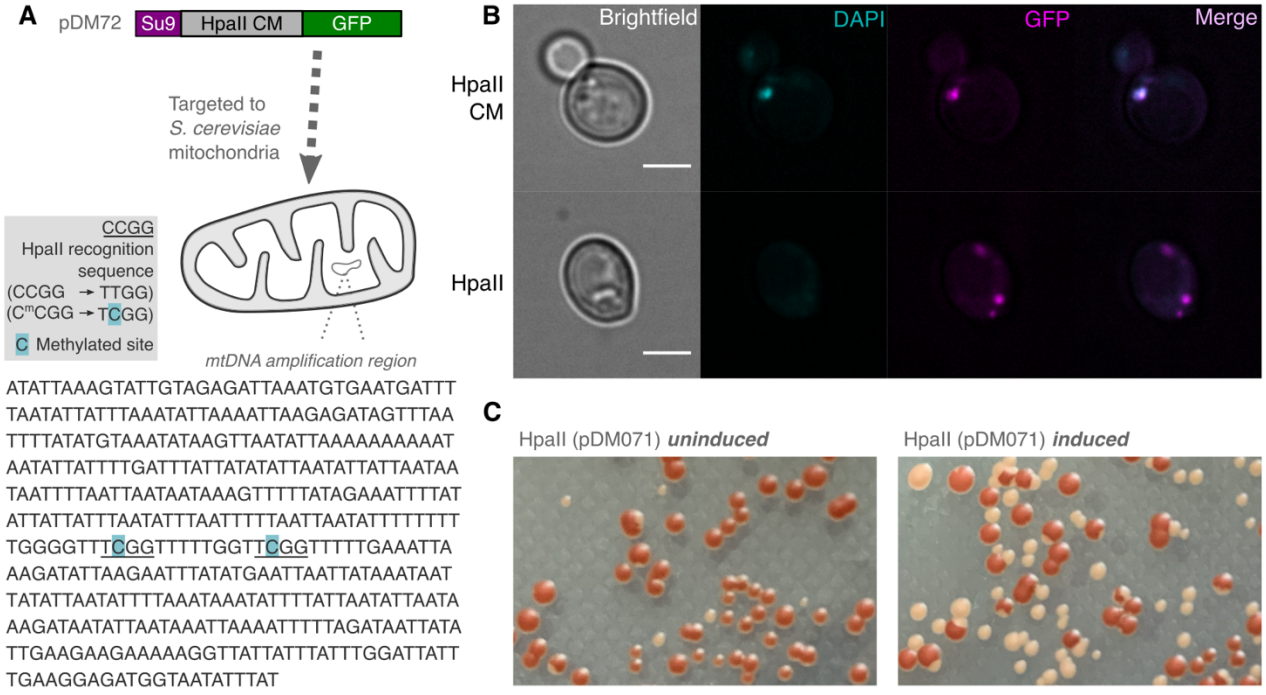


628

Figure 3



633 **Figure 4**



634

635

Figure legends

Fig. 1. The katablepharid K1 and K4 mitochondrial genomes encode two tandemly encoded restriction modification systems. A. The K4 mitochondrial genome encodes six ORFs (orange) with homologues in *L. marina* but with no similarity to any other eukaryote. Five ORFs are found in K4 but not in *L. marina* (red). These five genes include a putative GIY-YIG homing endonuclease, and two putative RM systems, each consisting of an endonuclease and a cytosine methyltransferase. Blue, protein coding genes; pink, RNA genes; black, introns. **B.** Variation in selfish elements detected in single amplified genomes and environmental DNA (eDNA). PCR followed by Sanger sequencing was performed to confirm integration of selfish element genes in K1, K3, and K4 single-cell amplified genomes (primer positions indicated by arrows). PCR of eDNA samples identified one product with high sequence identity (99.7%) to K4, and a shorter (2264 bp) unique product that is intermediate in length compared to K3 and K4. Red lines represent meta-transcriptome hits (95-100% identity) identified from the MATOU transcriptome database. Gene name abbreviations: *atp9*, ATP synthase 9 subunit; W, tryptophan tRNA; *HpaII*, HpaII endonuclease; *CM*, cytosine methylase; *MutH*, MutH endonuclease; *rns*, small subunit rRNA gene. Fig. 1B was generated using Clinker [61] and modified by hand.

Fig. 2. Phylogenetic reconstruction of concatenated genes encoded by katablepharid restriction-modification selfish elements, and comparison of codon frequencies between katablepharid and *Algibacter* complements. A. A concatenated phylogeny was reconstructed using sequences from K4 and 38 prokaryotic species containing tandemly encoded HpaII-CM and HpaII proteins. The concatenation resulted in an alignment length of 767 amino acid positions. Support values are posterior probabilities calculated using MrBayes v3.2.6 [50] and 1000 bootstrap replicates using RAxML v8.2.10 [51] and reported as MrBayes/RAxML. The MrBayes topology is shown. Species phyla are indicated as differently coloured branches as depicted inset. For individual trees of HpaII and HpaII-CM, see **Fig. S1**. **B.** Pairwise comparisons of sets of alternative codon frequencies for Kat4-HpaII-CM/HpaII, *Algibacter*-HpaII-CM/HpaII and the conserved protein-coding gene repertoire of the Kat4 mtDNA. Pairwise comparisons are shown in a grid. The key shows a grid with the corresponding amino acids. Results for Fisher exact tests comparing codon usage for each

amino acid are shown in tables between each pair. Asterisks denote significantly different codon usage, '-' indicates no significant difference in codon frequencies, and 'NA' indicates methionine and tryptophan, which were not tested as these amino acids are encoded by a single codon. Grids are placed on a grey circle between the three compared gene sets to identify the results of each pairwise comparison. Raw data available in **Supplementary File S3.**

Fig. 3. Heterologously expressed katablepharid HpaII-CM and HpaII are catalytically active. **A.** Bisulfite conversion to identify 5-methylcytosine modification by the katablepharid HpaII-CM. Schematic of bisulfite conversion protocol to assess 5-methylcytosine modifications. Plasmids were purified from *E. coli* Top10 and subjected to bisulfite conversion to convert cytosine to uracil (replaced with thymine during PCR), while 5-methylcytosines (5-mC) remain unaffected. 5-mC residues were detected within the amplification region when the katablepharid HpaII-CM was present on plasmid pDM040. Notably, each methylated site (indicated in blue) was located at CCGG, an HpaII recognition sequence (underlined). **B.** Transformation efficiency of *E. coli* strains when transformed with putative katablepharid HpaII-CM. Transformation efficiency of *E. coli* DH5α and Top10 strains when transformed with empty vector control (pACYC184) or vector containing the katablepharid putative methyltransferase coding sequence (pDM40). Experiments were performed from a minimum of three independent competent cell batches, and colony forming units (cfu) were enumerated and normalised to the positive control (pACYC184) within each batch. These data demonstrate that the katablepharid HpaII methyltransferase is toxic in *E. coli* DH5α (*mcrA*+), but not in Top10 (*mcrA*-). Error bars represent one standard deviation from the mean. **C.** Growth of *E. coli* Top10 cells with combinations of plasmids containing putative katablepharid methyltransferase (CM+), katablepharid HpaII (*Kat* HpaII) and Algibacter HpaII (*Alg* HpaII) genes, or the corresponding empty vectors ('no added CM' or 'no added REase' [restriction endonuclease]) . Duplicate cultures were grown for 8 h under Amp/Cm selection, induced with 0.0004% arabinose, at 37°C and growth was assessed by measuring OD_{595} at 5-minute intervals. The strain lacking the endonuclease showed typical *E. coli* growth, while addition of either the Algibacter (*Alg*) endonuclease or Katablepharid (*Kat*) endonuclease to the strain

lacking the methyltransferase caused toxicity. **D.** Addition of the katablepharid methyltransferase (CM+) rescued this toxicity to near control levels of growth (controls transposed from **C**). Error bars represent one standard deviation from the mean.

Fig. 4. Katablepharid HpaII endonuclease and methyltransferase induce petite mutants in *S. cerevisiae*. A. Bisulfite conversion to confirm targeting of a functional HpaII-CM to yeast mitochondria. Schematic of bisulfite conversion protocol to assess 5-methylcytosine modifications after induction of the katablepharid HpaII-CM from plasmid pDM072. 5-mC residues were detected within the amplification region of the *cox1* gene. Each methylated site (indicated in blue) was located at the HpaII recognition sequence (underlined). **B.** Evaluation of fluorescence for GFP-tagged HpaII-CM and HpaII, in conjunction with DAPI-labelled mtDNA. The HpaII-CM showed co-localisation with DAPI, while HpaII showed an absence of a DAPI focus, indicative of a lack of mtDNA, that is likely to be a product of endonuclease function and DNA degradation. Scale bar = 3 µm. **C.** HpaII expression causes petite formation. Formation of petite colonies (white) after the induction of HpaII (right), in comparison to an uninduced strain (left).

Supplementary figures

Figure S1. Phylogenetic reconstruction of HpaII (A) and HpaII-CM (B) encoded by katablepharid mitochondrial genome. Phylogenies were reconstructed using sequences from K4 and 38 prokaryotic species containing tandemly encoded HpaII and HpaII-CM proteins, resulting in alignments of 352 and 415 positions, respectively. Support values are posterior probabilities calculated using MrBayes v3.2.6 [50] and 1000 bootstrap replicates using RAxML v8.2.10 [51] and reported as MrBayes/RAxML. The MrBayes topology is shown.

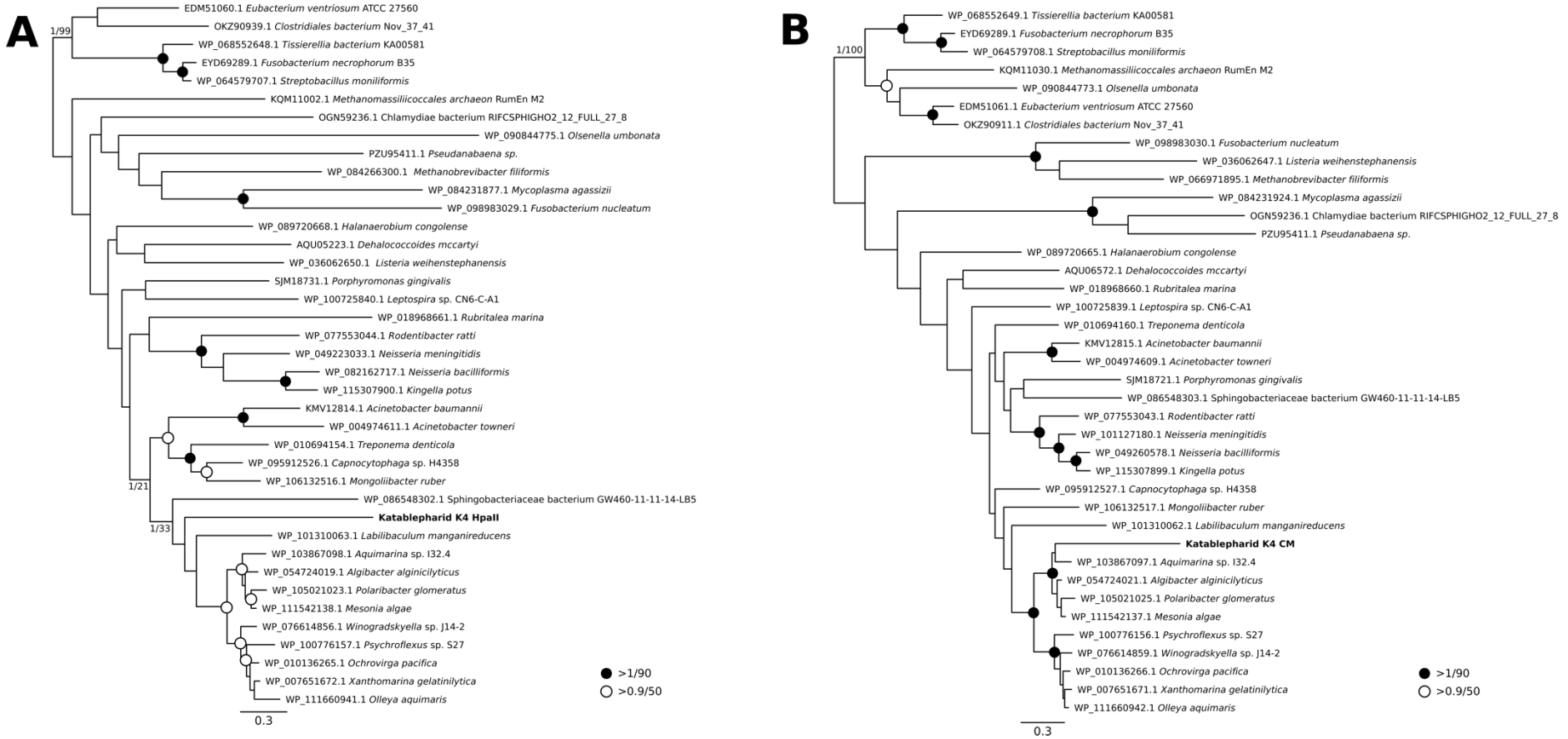


Figure S2. Katablepharid single amplified genomes contain no strong signal for bacterial contaminants. The contigs assigned to bacteria were low; as such, we have shown assignment only at the taxonomic level of 'Bacteria' and have not shown lower taxonomic divisions. Blob-plots were generated using BLOBTOOLS [62] for the three SAGs (K4, K1,and K3) that mapped to katablepharids using contigs >1000 bp. None of the contigs with best BLAST hits to bacteria were related to flavobacterial sequences.

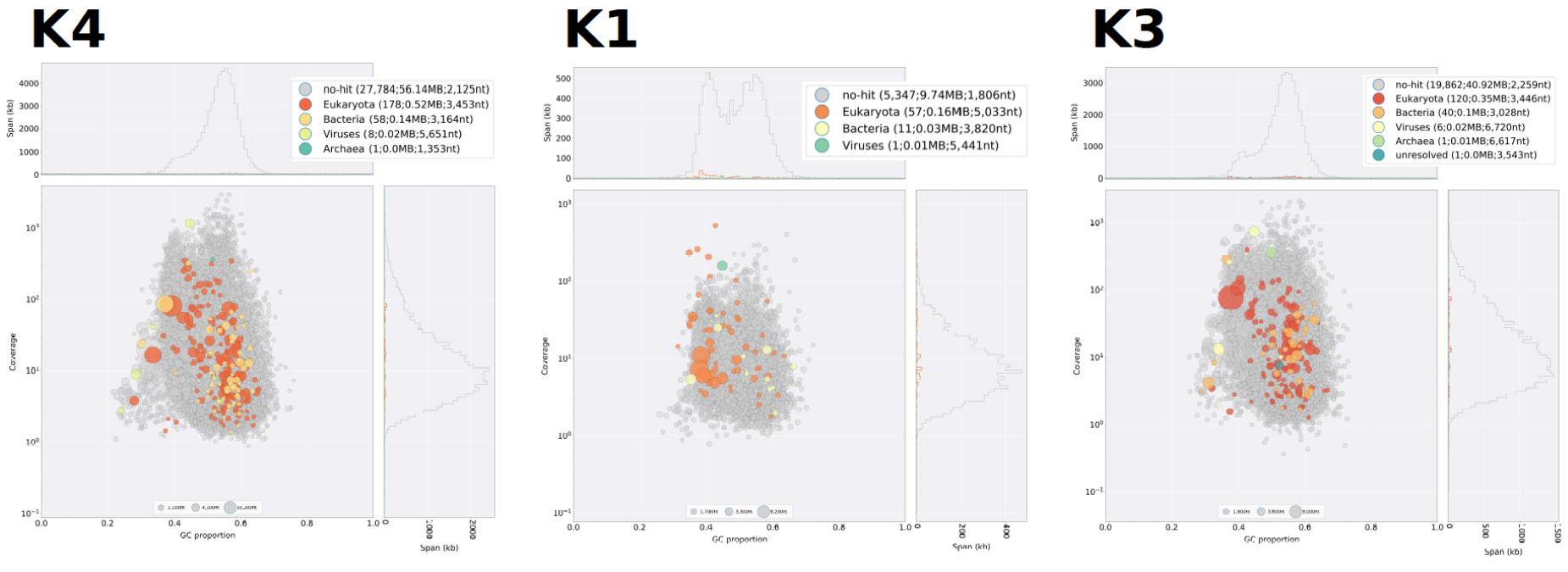


Figure S3. Growth assay of *E. coli* Top10 expressing putative katablepharid MutH-like endonuclease. Growth of *E. coli* Top10 cells with plasmid containing putative MutH-like endonuclease (MutH) genes, or the corresponding empty vector. Triplicate cultures were grown at 37°C for 8 h under Amp/Cm selection, induced with 0.1% arabinose, and growth was assessed by measuring OD_{595} at 5-minute intervals. This demonstrates that addition of the MutH-like endonuclease does not cause *E. coli* toxicity. Error bars represent 1 standard deviation from the mean.

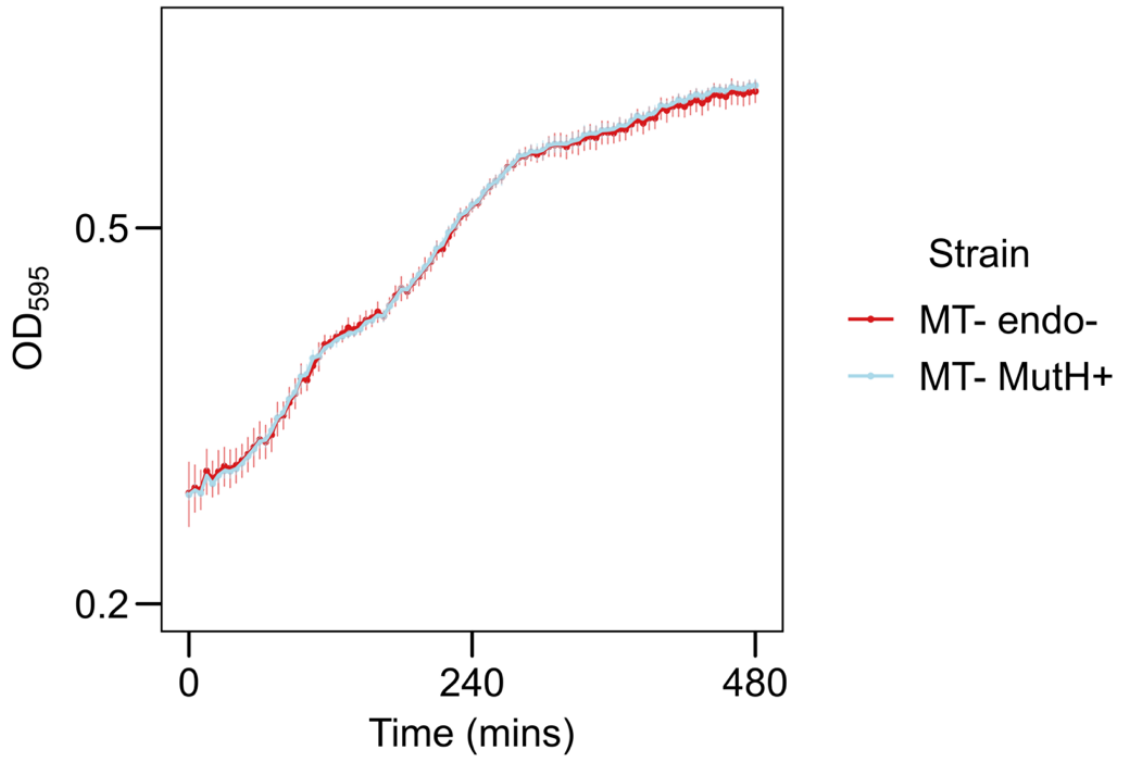


Figure S4. Perturbation of HpaII rescue by the katablepharid HpaII-CM after modifying the endonuclease constructs. Removal of the start codon of the ORF from the endonuclease pBAD expression vectors (leaving only the start codon encoded by the vector) resulted in the katablepharid HpaII-CM no longer offering protection against the katablepharid HpaII endonuclease (Kat HpaII) (**A**). However, the katablepharid HpaII-CM was still able to protect against the *Algibacter* HpaII endonuclease (Alg HpaII) (**B**), these results point towards a necessary concentration/function minimum requirement for rescue of katablepharid HpaII endonuclease.

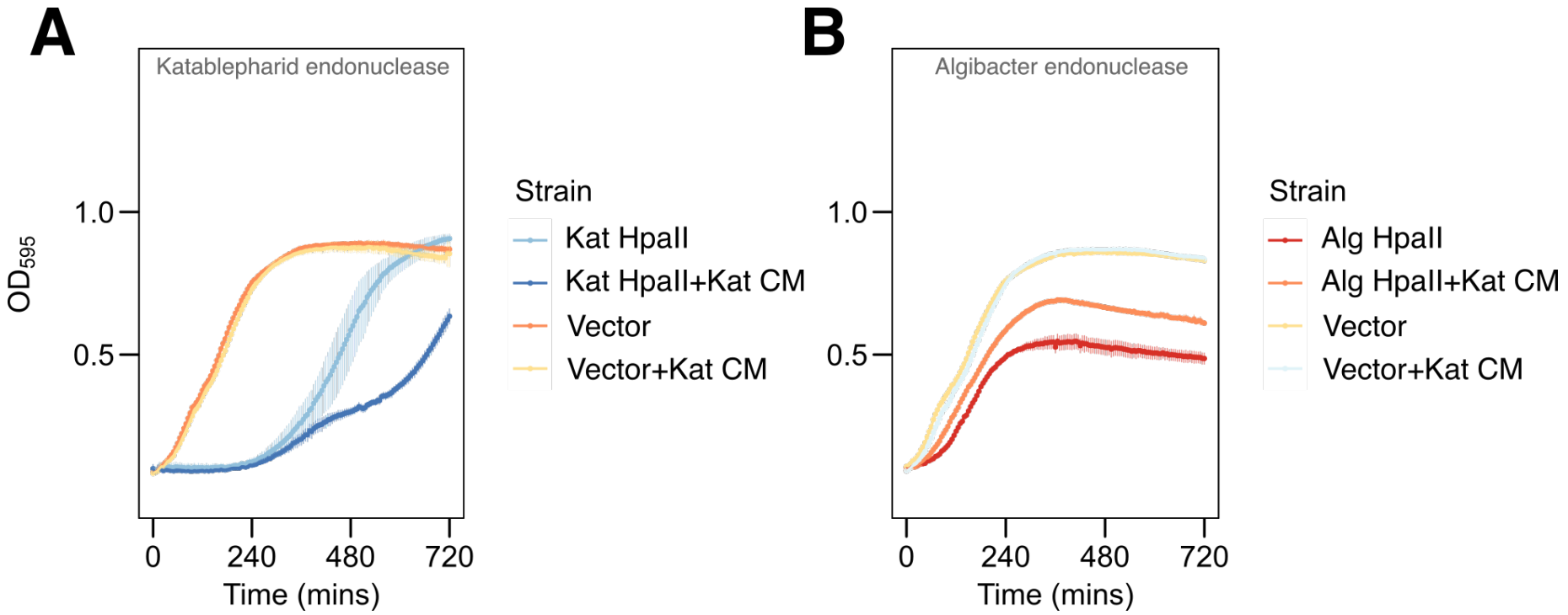


Table S1: Primers used in this study

Primer	Sequence (5'- 3')	Role
Atp9_F_upstream	CGTAGAAAATCAGAGGCGGC	Confirmation of RM selfish element in mitochondria
Atp9_F	ACGTAGGAGCAGGATTAGCAA	Confirmation of RM selfish element in mitochondria
Kat_F_2385	CGTTGGGATTAGTACCTTCCG	Confirmation of RM selfish element in mitochondria
Kat_F_3745	ACGCAAATCAGCAAGTGGTT	Confirmation of RM selfish element in mitochondria
Kat_R_1727	CGAGACTACCACGCCTCATA	Confirmation of RM selfish element in mitochondria
Kat_R_4022	TCACACCAACGACTAAAGCA	Confirmation of RM selfish element in mitochondria
rns_kat_mito_R2	CGTCCGCCTAAAACCTTTGT	Confirmation of RM selfish element in mitochondria
pACYC184_5mC_F	TAGTGGTGGTGAAATTTGATAGGATTATAA	Amplification of bisulfite-treated *E. coli* plasmid DNA (targeting CCGG sites)
pACYC184_5mC_R2	CAATTACCAATAACTACTACCAATAATACT	Amplification of bisulfite-treated *E. coli* plasmid DNA (targeting CCGG sites)
pACYC184_region2_5mC_F2	AAGATATGTAAAAGTATTATTGGTAGTAGT	Amplification of bisulfite-treated *E. coli* plasmid DNA (targeting GATC sites)
pACYC184_region2_5mC_R2	CCTACAACATCCAAAATAACAATACCAAAA	Amplification of bisulfite-treated *E. coli* plasmid DNA (targeting GATC sites)
cox1_bisulfite_F	ATATTAAAGTATTGTAGAGATTAAATGTGA	Amplification of bisulfite-treated *S. cerevisiae cox1* region
cox1_bisulfite_R	ATAAATATTACCATCTCCTTCAAATAATCC	Amplification of bisulfite-treated *S. cerevisiae cox1* region

Table S2: Plasmids used in this study. Note that all functions described are putative and constructs are codon optimised for expression in *E. coli*, unless otherwise stated.

Plasmid	Description
pDM027	pBAD HisA + Katablepharid HpaII endonuclease ORF
pDM034	pBAD HisA + *Algibacter* HpaII endonuclease ORF
pDM029	pBAD HisA + Katablepharid MutH endonuclease ORF
pDM040	pACYC184 + *Katablepharid* methyltransferase ORF
pDM041	pACYC184 + *Algibacter* methyltransferase ORF
pDM042	pACYC184 + Katablepharid MutH-like methyltransferase ORF
pDM071	pYES-mtGFP + Katablepharid HpaII endonuclease ORF (*S. cerevisiae* codon optimised)
pDM072	pYX223-mtGFP + Katablepharid HpaII methyltransferase ORF (*S. cerevisiae* codon optimised)
pDM076	pBAD HisA + Katablepharid HpaII endonuclease ORF, with ORF start codon omitted
pDM077	pBAD HisA + *Algibacter* HpaII endonuclease ORF, with ORF start codon omitted

Supplementary File S1: Data from BLASTN search against the Ocean Gene Atlas. Table of all BLASTN hits over 100 bp, with identity scores over 95%. Hits highlighted in grey align with the near-identical regions that flank the HpaII/HpaII-CM RM system; as such, these align with the 5' of both *hpaII CM* and *mutH CM*, so cannot be attributed to either gene.

Supplementary File S2: Sequence data of putative meiosis-associated genes identified in Katablepharid SAGs. Nucleotide and amino acid sequences for putative meiosis-associated proteins. Interruptions in open reading frames strongly suggest the presences of introns. Each protein was identified using the indicated hidden Markov model (HMM) and manually investigated. For each entry, the putative nucleotide sequence that could be confidently identified with BLAST is shown. The accession number for the parent contig of each sequence is provided and can be accessed here: https://figshare.com/articles/dataset/Single_Cell_Genomic_Assemblies/7352966

Supplementary File S3: Codon usage frequency data for proteins encoded by the Kat4 and *Algibacter* HpaII and HpaII-CM selfish elements, and the ancestral Kat4 mitochondrial proteins. Frequency of each amino acid codon for each of the Kat4-HpaII RM, the conserved protein-coding gene repertoire of the Kat4 mtDNA, and the *Algibacter*-HpaII RM. A Fisher exact test was used to compare codon usage frequencies for each pairwise comparison; p-values are displayed beneath each amino acid.