



HAL
open science

**Developmental timing of programmed DNA elimination
in *Paramecium tetraurelia* recapitulates germline
transposon evolutionary dynamics**

Coralie Zangarelli, Olivier Arnaiz, Mickaël Bourge, Kevin Gorrichon, Yan Jaszczyszyn, Nathalie Mathy, Loïc Escoriza, Mireille Bétermier, Vinciane Régnier

► **To cite this version:**

Coralie Zangarelli, Olivier Arnaiz, Mickaël Bourge, Kevin Gorrichon, Yan Jaszczyszyn, et al.. Developmental timing of programmed DNA elimination in *Paramecium tetraurelia* recapitulates germline transposon evolutionary dynamics. *Genome Research*, 2022, 2022, 10.1101/gr.277027.122 . hal-03840719v2

HAL Id: hal-03840719

<https://hal.science/hal-03840719v2>

Submitted on 3 Dec 2022 (v2), last revised 3 Jul 2023 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Developmental timing of programmed DNA elimination in *Paramecium tetraurelia* recapitulates**
2 **germline transposon evolutionary dynamics**

3
4 Coralie Zangarelli^{1,§}, Olivier Arnaiz^{1,§}, Mickaël Bourge¹, Kevin Gorrichon¹, Yan Jaszczyszyn¹,
5 Nathalie Mathy^{1,2}, Loïc Escoriza^{1,3}, Mireille Bétermier^{1*}, Vinciane Régnier^{1,4*}

6
7 ¹ Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198,
8 Gif-sur-Yvette cedex, France

9 ² Present address: Reproduction et Développement des Plantes UMR 5667, Ecole Normale Supérieure
10 de Lyon, 46 Allée d'Italie, 69364 Lyon Cedex 07, France.

11 ³ Present address: Laboratoire de génétique des hémopathies, Institut Universitaire du Cancer de
12 Toulouse, 1 avenue Irène Joliot-Curie, 31100 Toulouse, France.

13 ⁴ Université Paris Cité, UFR Sciences du Vivant, 75205 Paris Cedex 13, France

14 [§] These authors contributed equally to the work

15
16 *Correspondence: Mireille Bétermier, Vinciane Régnier

17
18 **Email:** mireille.betermier@i2bc.paris-saclay.fr; vinciane.regnier@i2bc.paris-saclay.fr

19
20 **Running title:** Timing of programmed DNA elimination in *Paramecium*

21
22
23
24
25

26 **Abstract**

27 With its nuclear dualism, the ciliate *Paramecium* constitutes an original model to study how host
28 genomes cope with transposable elements (TEs). *P. tetraurelia* harbors two germline micronuclei
29 (MIC) and a polyploid somatic macronucleus (MAC) that develops from the MIC at each sexual
30 cycle. Throughout evolution, the MIC genome has been continuously colonized by TEs and related
31 sequences that are removed from the somatic genome during MAC development. Whereas TE
32 elimination is generally imprecise, excision of ~45,000 TE-derived Internal Eliminated Sequences
33 (IESs) is precise, allowing for functional gene assembly. Programmed DNA elimination is
34 concomitant with genome amplification. It is guided by non-coding RNAs and repressive chromatin
35 marks. A subset of IESs is excised independently of this epigenetic control, raising the question of
36 how IESs are targeted for elimination. To gain insight into the determinants of IES excision, we
37 established the developmental timing of DNA elimination genome-wide by combining fluorescence-
38 assisted nuclear sorting with high-throughput sequencing. Essentially all IESs are excised within only
39 one endoreplication round (32C to 64C), while TEs are eliminated at a later stage. We show that DNA
40 elimination proceeds independently of replication. We defined four IES classes according to excision
41 timing. The earliest excised IESs tend to be independent of epigenetic factors, display strong sequence
42 signals at their ends and originate from the most ancient integration events. We conclude that old IESs
43 have been optimized during evolution for early and accurate excision, by acquiring stronger sequence
44 determinants and escaping epigenetic control.

45

46 **Introduction**

47 Transposable elements (TEs) have colonized the genomes of most living species and
48 constitute a significant fraction of extant genomes, from a few percent in yeast (Bleykasten-Grosshans
49 and Neugeglise 2011) to ~85% in some plant genomes (Bennetzen and Park 2018). TEs are often
50 considered as genomic parasites threatening host genome integrity, even though they can be a source
51 of genetic innovation (Cosby et al. 2019; Capy 2021). Host defense pathways counteract the
52 potentially detrimental effects of transposon invasion. In eukaryotes, small RNA (sRNA)-dependent
53 post-transcriptional and transcriptional silencing mechanisms inactivate TE expression and
54 transposition, both in germline and somatic cells (Ketting et al. 1999; Tabara et al. 1999; Zilberman et
55 al. 2003; Brennecke et al. 2007). TE transcriptional inactivation is associated with heterochromatin
56 formation, through DNA methylation and histone H3 methylation on lysine 9 (Deniz et al. 2019; Choi
57 and Lee 2020). Another epigenetic mark, H3K27me3, also contributes to TE silencing in several
58 species (Dél  ris et al. 2021).

59 Because of their germline-soma nuclear dualism (Prescott 1994; Cheng et al. 2020), ciliates
60 are original unicellular eukaryotic models to study the dynamics of TEs within genomes, both at the
61 developmental and evolutionary time-scales (Arnaiz et al. 2012; Hamilton et al. 2016; Kapusta et al.
62 2017; Sellis et al. 2021). *Paramecium* species harbor one to four transcriptionally silent diploid
63 germline micronuclei (MIC) (G  rtz 1988) that coexist with a polyploid somatic macronucleus (MAC)
64 responsible for gene expression. During sexual processes (conjugation of compatible reactive partners
65 or a self-fertilization process called autogamy), the MICs undergo meiosis and transmit the germline
66 genome to the diploid zygotic nucleus through fertilization and karyogamy (B  termier and Duharcourt
67 2014). In the meantime the old MAC splits into ~30 fragments that continue to ensure gene expression
68 while new MICs and MACs differentiate from division products of the zygotic nucleus. The formation
69 of a functional new MAC is essential to take over gene expression once old MAC fragments have
70 disappeared from the cell. New MAC development covers two cell cycles after the zygotic nucleus is
71 formed. During this period, massive genome amplification takes place within each developing MAC
72 (also called anlagen) to reach the final endoduplication level of mature MACs (~800C to 1600C in *P.*
73 *tetraurelia*) (Preer 1976). Concomitantly with genome amplification, programmed DNA elimination

74 (PDE) removes ~30% of germline DNA from the new MAC genome, going from 98 - 151 Mbp
75 haploid genome size in the MIC to 72 - 75 Mbp in the mature MAC (Aury et al. 2006; Guérin et al.
76 2017; Sellis et al. 2021). Because eliminated DNA includes TEs and related sequences (Arnaiz et al.
77 2012; Guérin et al. 2017; Sellis et al. 2021), PDE in *Paramecium*, as in other ciliates, can be viewed as
78 an extreme mechanism to inactivate TEs in the somatic genome.

79 Two types of germline sequences, referred to as "MIC-limited" DNA, are removed during
80 PDE in *Paramecium* (Bétermier and Duharcourt 2014). At least 25% of the MIC genome, including
81 DNA repeats (TEs, minisatellites), are eliminated imprecisely, alternatively leading to chromosome
82 fragmentation (with *de novo* telomere addition to heterogeneous new MAC chromosome ends) or
83 intrachromosomal deletions between variable boundaries (Baroin et al. 1987; Le Mouëll et al. 2003;
84 Guérin et al. 2017). In contrast, ~45,000 Internal Eliminated Sequences (IESs) scattered throughout
85 the germline genome (including inside coding sequences) are excised precisely, allowing assembly of
86 functional open reading frames (Arnaiz et al. 2012). *Paramecium* IESs are mostly short (93% <150
87 bp) non-coding sequences, with a damped sinusoidal size distribution extending from 25 bp to a few
88 kbp. They are consistently flanked by two TA dinucleotides, one on each side, and leave a single TA
89 on MAC chromosomes upon excision. Two independent studies, the first relying on the analysis of
90 paralogous gene quartets originating from successive whole genome duplications in a single species,
91 *P. tetraurelia* (Arnaiz et al. 2012), the other on phylogenetic analyses across 9 *Paramecium* species
92 (Sellis et al. 2021), have made it possible to date ~40% of *P. tetraurelia* IES insertions and define
93 groups of old, intermediate and young IESs according to their evolutionary age. The oldest IESs,
94 thought to have colonized the germline genome before divergence of *P. caudatum* and the *P. aurelia*
95 clade, tend to be very short (26 to 30 bp) (Sellis et al. 2021). Several families of larger and younger
96 IESs, some sharing homology with known *Paramecium* TEs, appear to have been mobile recently at
97 the time-scale of *Paramecium* evolution: intermediate IESs were acquired after the divergence of *P.*
98 *caudatum*, young IESs were gained after the burst of *P. aurelia* speciation. This is consistent with
99 IESs being relics of ancestral TEs that have decayed during evolution through reduction in size and
100 loss of coding capacity, while remaining under selection for precise excision from the MAC
101 (Klobutcher and Herrick 1997; Dubois et al. 2012).

102 IES excision occurs through a “cut-and-repair” mechanism involving double-strand DNA
103 cleavage around each flanking TA (Gratias and Bétermier 2003), followed by excision site closure
104 through precise Non-Homologous End Joining (NHEJ) (Kapusta et al. 2011; Bétermier et al. 2014).
105 Several components of the core IES excision machinery are known. The PiggyMac (Pgm)
106 endonuclease, a catalytically active domesticated transposase (Baudry et al. 2009; Dubois et al. 2012),
107 and its five PiggyMac-like partners, PgmL1 to PgmL5 (Bischerour et al. 2018), are essential for the
108 introduction of DNA double-strand breaks at IES ends. In the absence of Pgm, all IESs are retained in
109 the anlagen and most imprecise DNA elimination is also impaired, except for ~3 Mbp of germline
110 sequences, the elimination of which seems to be Pgm-independent (Guérin et al. 2017). A specialized
111 NHEJ factor, the Ku70/Ku80c (Ku) heterodimer also appears to be an essential component of the core
112 endonuclease machinery: Ku is able to interact with Pgm, tethers it in the anlagen and licenses DNA
113 cleavage at IES ends (Marmignon et al. 2014; Abello et al. 2020; Bétermier et al. 2020).

114 *Paramecium* IES ends display a weak consensus (5' TAYAGTNR 3'), which includes the
115 palindromic flanking TA dinucleotide conserved at each boundary (Arnaiz et al. 2012). This
116 consensus defines an internal inverted repeat at IES ends but is too poorly conserved to serve as a
117 specific recognition sequence for the endonuclease. Additional epigenetic factors, including non-
118 coding RNAs and histone modifications, control the recognition of eliminated DNA by the core
119 machinery (Chalker et al. 2013; Bétermier and Duharcourt 2014; Allen and Nowacki 2020).
120 According to the “scanning” model, sRNAs processed from meiotic MIC transcripts by Dicer-like
121 proteins Dcl2 and Dcl3 (called “scnRNAs”) are subtracted against old MAC sequences, resulting in
122 the selection of a sub-population of scnRNAs covering the MIC-limited fraction of the germline
123 genome (Lepère et al. 2008; Lepere et al. 2009). MIC-limited scnRNAs are thought to target
124 elimination of their homologous sequences by pairing with TFIIS4-dependent non-coding nascent
125 transcripts in the anlagen (Maliszewska-Olejniczak et al. 2015), thereby triggering H3K9 and K27
126 trimethylation by the PRC2 complex containing the histone methyltransferase Ez11 (Lhuillier-Akakpo
127 et al. 2014; Frapporti et al. 2019; Miró-Pina et al. 2022; Wang et al. 2022). The H3K9me3 and
128 H3K27me3 heterochromatin marks are required for the elimination of TEs and ~70% of IESs
129 (Lhuillier-Akakpo et al. 2014; Guérin et al. 2017). A second population of sRNAs (called iesRNAs),

130 produced by the Dcl5 protein from excised IES transcripts, was proposed to further assist IES excision
131 (Sandoval et al. 2014; Allen et al. 2017). Both types of sRNAs appear to act synergistically. Indeed,
132 while *DCL2/3* or *DCL5* knockdowns (KD) each impair excision of only a small fraction of IESs (~7%
133 in a *DCL2/3* KD, ~5% in a *DCL5* KD) (Lhuillier-Akakpo et al. 2014; Sandoval et al. 2014), a triple
134 *DCL2/3/5* KD inhibits excision of ~50 % of IESs coinciding with the set of TFIIS4-dependent IESs
135 (Swart et al. 2017).

136 While our knowledge of the molecular mechanisms involved in the epigenetic control and
137 catalysis of PDE in *P. tetraurelia* has increased over the past decade, little is known about the relative
138 timing of DNA replication and PDE during MAC development. Molecular data obtained for a handful
139 of IESs suggested that excision starts following several endoreplication rounds in the anlagen
140 (Bétermier et al. 2000). In the present study, we have investigated at the genome-wide level the
141 elimination timing of all 45,000 IESs and other MIC-limited sequences, including TEs. To follow the
142 progression of IES excision during MAC development, we monitored for each IES the fraction of
143 excised molecules that were present in purified anlagen at each developmental stage, which we have
144 referred to as the excision score (ES). This allowed us to establish the timing of PDE during MAC
145 development, address whether a mechanistic link exists between IES excision and DNA replication,
146 and examine whether the temporal and epigenetic control of PDE may be related to the evolutionary
147 age of eliminated DNA.

148

149 **Results**

150 **A Fluorescence-Activated Nuclear Sorting (FANS) strategy to purify new MACs**

151 Because old MAC fragments containing the rearranged genome are present in *Paramecium*
152 cells throughout the sexual cycle, we set up a protocol to selectively purify developing new MACs
153 during an autogamy time-course (tc). We adapted a published flow cytometry procedure that was
154 initially designed to sort anlagen from old MAC fragments at a late developmental stage, when the
155 two types of nuclei can clearly be distinguished based on their size and DNA content (Guérin et al.
156 2017). Because at early stages (DEV1 and DEV2, see Methods), anlagen and old MAC fragments
157 have similar sizes (Fig. 1A), we selectively labeled the new MACs using a specific α -PgmL1 antibody

158 raised against a component of the IES excision machinery (Bischerour et al. 2018). We first confirmed
159 that immunofluorescence staining of whole cells yielded a strong and specific signal in the anlagen
160 throughout DEV1 to DEV4 (Fig. 1A and Supplemental Fig. S1A-C), which corresponds to the time-
161 window (T3 to T30) when programmed double-strand breaks are detected at IES boundaries (Gratias
162 et al. 2008; Baudry et al. 2009). Using the α -PgmL1 antibody to label unfixed nuclei harvested at
163 DEV1 and DEV2 during an autogamy time-course of wild-type (wt) cells, we confirmed that PgmL1
164 labeling can be used to separate anlagen from old MAC fragments using flow cytometry (Fig. 1B and
165 Supplemental Fig. S1D,E).

166

167 **Most IES excision takes place within one round of replication**

168 We used autogamy to purify new MACs by FANS at different DEV stages despite the
169 asynchrony of this sexual process (Berger 1986), because it allows us to collect large amounts of
170 material. The distribution of propidium iodide (PI) fluorescence intensities revealed a series of three
171 discrete peaks from DEV1 to DEV3 for PgmL1-labelled new MACs (Fig. 2A). This is indicative of
172 the presence of nuclear populations with a defined DNA content. Previously published work suggested
173 that at least 4 discontinuous peaks of DNA synthesis, corresponding to ~ 5 doublings of DNA content,
174 take place in anlagen before the first cell fission in *P. tetraurelia*, while 4.5 additional doublings occur
175 with a more continuous pattern during the second cell cycle (Berger 1973). We therefore made the
176 reasonable assumption that each peak observed in flow cytometry corresponds to one whole genome
177 doubling following a pulse of genome replication, and focused on these populations to draw the
178 sorting gates for further purification. We calculated the DNA content for each peak (C-value in Mbp,
179 see Methods, Supplemental Fig. S2 and Table S1) and further defined the corresponding amplification
180 level of the genome (C-level), using an approximate 1C-value of 100 Mbp for the unarranged *P.*
181 *tetraurelia* MIC genome (Guérin et al. 2017; Sellis et al. 2021). We attributed the closest power of 2 to
182 each resulting amplification level and defined an estimated C-level of $\sim 32C$ (DEV1/DEV2), $\sim 64C$
183 (DEV2/DEV3) and $\sim 128C$ (DEV3) for each population (tc4 in Supplemental Table S1). At DEV4,
184 which is the final stage where PgmL1 staining can be detected, we observed an enlargement of the
185 $\sim 128C$ peak, indicative of a mixed population with a more variable amount of DNA (see Discussion).

186 We further sorted the populations of nuclei issued from each peak (Fig. 2A) and extracted
187 their DNA for deep sequencing (tc4 in Supplemental Table S2). Thanks to the absence of old MAC
188 contamination (See Methods and Supplemental Fig. S3A-C), molecules lacking an inserted IES
189 (designated IES⁻) only correspond to *de novo* excision junctions. Therefore, the power of the FANS
190 procedure allows us for the first time to calculate a real excision score (ES) for each of the 45,000
191 IESs (Fig. 2B), which varies from 0 (no excision) to 1 (complete excision). At DEV1 ~32C, few IESs
192 have been excised, with a median ES value of 0.15. The median ES rises to 1 at DEV3 ~64C,
193 indicating that nearly all IESs are excised within one round of replication. To investigate whether the
194 5th endoreplication round itself is mandatory for DNA elimination, we performed a replicate time-
195 course experiment in which we treated autogamous cells with aphidicolin, a specific inhibitor of
196 eukaryotic replicative DNA polymerases (Byrnes 1984; Cheng and Kuchta 1993), after they reached
197 DEV1 ~32C (Fig. 2C). Comparison of the flow cytometry profiles confirmed that the new MACs of
198 control cells (+DMSO) have undergone their 5th replication round at DEV3, while those of
199 aphidicolin-treated cells are blocked at ~32C. We further sorted anlagen from the DEV1, DEV3
200 DMSO and DEV3 Aphi samples for DNA sequencing (Supplemental Table S2). For the control
201 replicate, we confirmed that most IES excision is completed within one round of replication, between
202 ~32C and ~64C (median ES at DEV3 ~64C is 0.99; Fig. 2D). For aphidicolin-treated anlagen, the
203 median ES is 0.98, indicating that inhibiting the 5th endoreplication round does not impair IES
204 excision.

205

206 **Imprecise elimination is delayed relative to IES excision**

207 To strengthen our analysis of the timing of DNA elimination, we included sorted samples
208 from 4 additional replicate time-course experiments (Supplemental Fig. S4A-C and Tables S1, S2).
209 The resulting ES distributions confirm our conclusion that IES excision takes place between DEV1
210 ~32C and DEV3 ~64C (Fig. 3A). We used the same sequencing data (Supplemental Table S2) to
211 study the timing of imprecise DNA elimination during MAC development. Because this process yields
212 heterogeneous MAC junctions, preventing us from calculating an ES, we analyzed sequencing data by
213 read coverage (Fig. 3B). Using this procedure, we confirmed that the sequencing coverage drops

214 between DEV1 ~32C and DEV3 ~64C for IESs, consistent with the excision profile obtained by ES
215 calculation (Fig. 3A). After this validation, we analyzed TE coverage as a proxy for imprecise DNA
216 elimination. We observed a delayed decrease relative to IES coverage, with a drop starting at DEV3
217 ~64C. Analysis of the percentage of coverage of the whole MIC genome revealed a similar decrease
218 (from 97 to 90%) between DEV3 ~64C and DEV3 ~128C, which likely corresponds to the elimination
219 of MIC-specific DNA. During DNA elimination, however, TE sequences may be found in two forms:
220 non-excised intrachromosomal molecules and not yet degraded extrachromosomal elimination
221 products. Because sequence coverage analysis cannot discriminate between these two forms, TE
222 elimination may have started before the drop of sequence coverage. We therefore used another marker
223 of imprecise elimination: the formation of *de novo* telomeric ends that accompanies the removal of
224 TE-containing MIC-specific sequences (Le Mouël et al. 2003). We observed that telomeric reads only
225 increase at DEV3 ~64C (Fig. 3C), supporting the idea that imprecise elimination does not begin before
226 DEV3. Of note, the majority of telomere addition sites are localized at more than 100-nt distance from
227 IES boundaries, confirming that they are not related to precise IES elimination. We also noticed that
228 the whole MIC genome coverage at DEV4 ~128C is still higher than the genome coverage of
229 fragments (which harbor a fully rearranged genome), indicating that imprecise elimination is not
230 totally completed in the new MAC at this stage.

231

232 **Genome wide detection of transient IES-IES junctions**

233 We took advantage of the purity of FANS-sorted anlagen to increase our ability to detect
234 transient DNA molecules produced during IES excision. Based on a few Southern blot experiments,
235 IESs were proposed to be excised as linear molecules and subsequent formation of closed DNA circles
236 was documented for a few long IESs (Gratias and Bétermier 2001). More recently, excised IESs were
237 proposed to concatenate through the NHEJ pathway into end-to-end joined circular molecules that are
238 used as substrates for transcription and Dcl5-dependent production of iesRNAs, before being
239 eventually degraded (Allen et al. 2017). The existence of multi-IES concatemers, however, was only
240 supported by the sequencing of reverse-transcribed RNA molecules, and direct evidence for

241 concatemerized DNA molecules was still lacking. We therefore developed a new bioinformatic
242 method to quantify IES excision products from high-throughput DNA sequencing data.

243 Among the three expected types of excision products (linear molecules, single-IES circles and
244 multi-IES concatemers), only IES-IES junctions from circles or concatemers were analyzed. Indeed,
245 the sequencing reads that map within IESs cannot be used to unambiguously count linear excised
246 molecules, because they do not discriminate between intrachromosomal (not excised) or
247 extrachromosomal (excised) IES forms. We confirmed that new MACs indeed contain DNA
248 molecules corresponding to single-IES circles and multi-IES concatemers (Fig. 3D and Supplemental
249 Fig. S5A). Because the normalized count of IES-IES junctions is maximal at DEV2 ~32C (Fig. 3D),
250 the stage at which the ES increases (Fig. 3A), we infer that IES-IES junctions may be formed
251 concomitantly with MAC junctions but are still detected at ~64C DEV3, when IES excision is
252 completed. This confirms, at the genome-wide level, that excised IES products are not degraded
253 immediately and persist in the new MACs (Bétermier et al. 2000). Our data also reveal that the vast
254 majority of IESs (97.2% considering all datasets and 86% at DEV2 ~32C) are involved in the
255 formation of IES-IES junctions (Fig. 3D). Based on read counts, single-IES circles represent fewer
256 than 2% of excised IES junctions (Supplemental Fig. S5B), indicating that concatemers are the major
257 products of IES-end joining following excision. This can be explained by the size distribution of IESs,
258 93% being shorter than 150 bp (Armaiz et al. 2012), a size corresponding to the persistence length of
259 double-stranded DNA, below which self-circularization is inefficient (Schleif 1992; Bates et al. 2013).
260 Consistently, we find that the size distribution of single-IES circles is centered around 200 bp with a
261 sharp drop for IESs shorter than 150 bp (Supplemental Fig. S5C). Our sequencing data do not allow us
262 to determine the size range of IES concatemers but previous experimental observations have indicated
263 that concatemeric and single-IES circles have the same size range (above 200 bp) (Gratias and
264 Bétermier 2001; Allen et al. 2017), and support our conclusion that only the longest IESs can self-
265 circularize.

266

267 **Sequential timing of excision is associated with specific IES features**

268 Previously published molecular data suggested that not all IESs are excised at exactly the
269 same time (Gratias and Bétermier 2001). To gain deeper insight into the differential timing of IES
270 excision, we used the ES values obtained for the ~45,000 IESs across all samples to group IESs into 4
271 clusters, according to their excision timing ("very early", "early", "intermediate", "late"; Fig. 4A;
272 Supplemental Fig. S6A and Table S3). Very early IESs are almost all excised at DEV2 ~32C, while
273 excision of most late IESs takes place between DEV2 ~64C and DEV3 ~64C. Detection of IES-IES
274 junctions follows the same excision timing: IESs from the very early and early clusters contribute to
275 the majority of junctions detected at the earliest developmental stages, while IESs from the
276 intermediate and late clusters become dominant at late developmental stages (Supplemental Fig. S6B).
277 It has been previously observed that the excision machinery sometimes generates different types of
278 errors caused by the use of misplaced alternative TA boundaries (Supplemental Fig. S7A) (Duret et al.
279 2008; Bischerour et al. 2018). At the stage when all IESs are completely excised (DEV4 ~128C), we
280 observe 7-fold fewer excision errors for very early relative to late excised IESs (Fig. 4B), indicating
281 that very early IESs are much less error-prone. Of note, the maximum of excision errors during the
282 excision time-course never exceeds the error level observed in old MAC fragments (Supplemental Fig.
283 S7B,C).

284 With regard to genomic location, we found that late IESs are under-represented in genes,
285 particularly in coding sequences (CDS) versus introns, while the inverse trend is observed for very
286 early and early IESs (Supplemental Fig. S8A). We also observed a strong enrichment of late excised
287 IESs and a depletion of very early and early IESs at the extremities of MAC scaffolds, which is
288 consistent with these regions being gene-poor (Supplemental Fig. S8B). Under-representation of late
289 excised IESs within genes might be explained by a selective pressure for accurate excision to avoid
290 the formation of non-functional ORFs.

291 As for IES intrinsic properties, we detected an impressive size bias between IESs from the
292 different clusters, with very early excised IESs tending to be much shorter than expected from the
293 global IES size distribution. In contrast, short IESs are under-represented among late IESs (Fig. 4C
294 and Supplemental Fig. S9). We then examined whether IESs have different sequence properties at

295 their ends depending on the cluster to which they belong. Because the consensus of IES ends varies at
296 positions 3, 4 and 5 (position 1 being the T from the TA boundary) as a function of IES length (Swart
297 et al. 2014), we compared the sequence logos of IES ends in the different clusters according to IES
298 size category (Fig. 4D and Supplemental Fig. S10A,B). For 25 to 33-bp IESs, we found an over-
299 representation of the TATAG boundary among very early IESs compared to late IESs, with a
300 significant increase of G frequency at the 5th base position (62% vs 35% for very early compared to
301 late IESs). For 42 to 140-bp IESs, we observed an even stronger sequence bias with an over-
302 representation of the TACAG boundary among very early IESs, the increase of the C frequency at the
303 third position being highly significant (77% vs 30% for very early vs late IESs, respectively). We
304 conclude that very early IESs shorter than 140 bp tend to exhibit a stronger nucleotide sequence signal
305 at their ends than late excised IESs. No significant sequence difference between very early and late
306 IESs was observed for longer IESs (>140 bp).

307 We further studied the link between excision timing and dependence upon known factors
308 involved in the epigenetic control of IES excision (Fig. 4E,F and Supplemental Fig. S11). We found
309 an under-representation of the very early excised cluster amongst the subset of IESs whose excision
310 depends on the deposition of H3K9me3 and H3K27me3 marks (*i.e.* IESs retained in an *EZLI* KD)
311 (Lhuillier-Akakpo et al. 2014). A similar bias was observed among IESs depending on the production
312 of TFIIIS4-dependent transcripts from the anlagen (Maliszewska-Olejniczak et al. 2015) and was
313 exacerbated for sRNA-dependent IESs (retained in *DCL2/3* or *DCL5* KDs) (Sandoval et al. 2014;
314 Lhuillier-Akakpo et al. 2014). In the *DCL* RNAi datasets, IESs from the very early cluster are totally
315 absent, while IESs from the intermediate and the late clusters are over-represented. In contrast, very
316 early excised IESs are strongly enriched (~60%) among the 12,414 IESs that are excised
317 independently of the above factors ("excision complex only"). Considering the overlap between IES
318 dependencies (Fig. 4E), our data indicate that IESs depending on known heterochromatin-targeting
319 factors tend to take longer to be excised during MAC development. Consistent with late IESs being
320 error-prone (Fig. 4B), we observed more errors for *Dcl2/3*- or *Ez11*-dependent IESs than for IESs
321 depending on the "excision complex only" (Supplemental Fig. S7D).

322 Finally, we examined the relationship between IES evolutionary age (Sellis et al. 2021) and
323 excision timing (Fig. 4G). Our data indicate that old IESs that invaded the *Paramecium* genome before
324 the divergence of the *P. caudatum* and *P. aurelia* lineages tend to be precociously excised.
325 Reciprocally, we observed that the younger the IESs, the later their excision during MAC
326 development.

327

328 **Identification of new IESs in MIC-limited regions**

329 The presence of IESs nested in MIC-limited regions was reported previously but only a few
330 examples have been described (Mayer et al. 1998; Duharcourt et al. 1998; Mayer and Forney 1999; Le
331 Mouël et al. 2003). We took advantage of the sequencing data we obtained during the course of MAC
332 development to pinpoint precise excision events within late eliminated regions, therefore identifying
333 new *bona fide* IESs (see Supplemental Methods). Their excision could be transiently observed before
334 complete elimination of the surrounding DNA. We could identify a set of 167 “buried” IESs localized
335 in imprecisely eliminated regions and 226 “internal” IESs located inside IESs from the reference set
336 (Supplemental Fig. S12A,B and Tables S4, S5). We found that buried IESs are strongly biased
337 towards short sequences while internal IESs present no major difference in size compared to the
338 reference IESs (Supplemental Fig. S12C).

339 In order to assess whether these newly identified IESs depend on heterochromatin marks for
340 excision, we analyzed their retention in Ez11-depleted cells (Supplemental Fig. S12D and Table S6).
341 As previously published (Denby Wilkes et al. 2016), we calculated their retention scores (IRS: IES
342 retention score), varying from 0 for no retention to 1 for full retention. We found two contrasting
343 situations for internal IESs: 26% are not affected in Ez11-depleted relative to control cells (IRS ~0)
344 while 20% are strongly retained (IRS ~1). We also noticed that retained IESs are much longer than the
345 unaffected ones (boxplot in Supplemental Fig. S12D). A similar size bias was reported for the Ez11-
346 dependent IESs from the reference set (Lhuillier-Akakpo et al. 2014). For their related encompassing
347 IESs (n=223), we observed that 94% are significantly retained in Ez11-depleted cells, consistent with
348 their late excision timing. Our results suggest that internal IESs exhibit similar characteristics to those
349 of the reference IES set in terms of length and epigenetic control, and therefore might share the same

350 evolutionary history. With regard to the excision mechanism, molecular data indicate that IESs can be
351 excised while retaining a nested IES (Bétermier et al. 2000; Gratias and Bétermier 2001). This
352 suggests that excision of internal IESs is not a systematic prerequisite for the elimination of their
353 encompassing IES, reminiscent of the sequential splicing of introns inserted within introns (Hafez and
354 Hausner 2015). Thus, the existence of internal IESs adds to the list of features shared by IESs and
355 introns (Arnaiz et al. 2012; Sellis et al. 2021).

356 In contrast to internal IESs, we found that most buried IESs are independent of Ez11-mediated
357 heterochromatin marks for their excision (Supplemental Fig. S12D). Moreover, we noticed that the
358 most independent are the shortest, with a breakpoint size of 33 nt (Supplemental Fig. S12E). The
359 finding of buried IESs confirms that IESs are scattered all along MIC chromosomes, including MIC-
360 limited regions, as previously hypothesized (Sellis et al. 2021). The properties of buried IESs,
361 however, raise the question of their origin. Most IESs are derived from TEs, but a previous report
362 showed that genomic fragments can be co-opted to become IESs (Singh et al. 2014). Even though
363 buried IESs have no stronger sequence end logos than the set of all IESs (Supplemental Fig. S12F), we
364 speculate that buried IESs are excision-prone genomic fragments recognized by the excision
365 machinery independently of histone mark deposition. Why these genomic fragments are excised as
366 IESs remains an open question.

367

368 **Discussion**

369 **Developmental timing of sequential DNA elimination**

370 The present study aimed at unravelling the links between two intertwined DNA-driven
371 mechanisms underlying somatic nuclear differentiation in *Paramecium*: genome amplification and
372 programmed DNA elimination. Setting up the FANS procedure allowed us to demonstrate that
373 genome amplification during MAC development is an endocycling process, defined as alternating S
374 and G phases without mitosis (Lilly and Duronio 2005). In the time-window during which PDE takes
375 place, we identified three peaks representing discrete new MAC populations differing in their DNA
376 content. The estimated C-levels of the first two peaks (~32C, ~64C) are consistent with their resulting
377 from successive whole-genome doublings. The range of C-levels obtained for the third peak fits less

378 well with ~128C, which can be explained by ongoing massive DNA elimination between ~64C and
379 ~128C causing variability in the actual 1C-value of the anlagen.

380 We determined at an unprecedented resolution the timing of IES excision and imprecise DNA
381 elimination genome-wide, across successive endoreplication cycles (Fig. 5A). Our data show that
382 DNA elimination is an ordered process. We found that most IESs are excised between DEV1 ~32C
383 and DEV3 ~64C under standard conditions (see Methods), while imprecise elimination only starts at
384 DEV3 ~64C. We established four classes of IESs according to their excision timing (Fig. 4A). We
385 also demonstrated that the progression of IES elimination, once it has started, is independent of
386 replication, suggesting that the excision machinery is recruited to its chromatin targets independently
387 of replication fork passage.

388 We observed that little IES excision has already taken place at the earliest stage of our study
389 (median ES = 0.15 at DEV1 32C), suggesting that the onset of PDE is controlled during MAC
390 development. Given that *Paramecium* IESs are mostly intragenic (Arnaiz et al. 2012), starting their
391 excision after a few endoreplication rounds have taken place may have been advantageous to limit the
392 detrimental effects of excision errors on functional gene assembly. The temporal control of PDE may
393 be explained by the expression profile of genes encoding components of the core excision machinery,
394 many of which (*e.g.* *PGM*, *PGMLs*, *KU80C*) are not expressed during early autogamy stages and
395 reach their maximum transcription level at DEV2 (Arnaiz et al. 2017). In addition, as previously
396 suggested (Bétermier et al. 2000), the 3 to 4 endoreplication rounds preceding IES excision might also
397 contribute to chromatin remodeling, to provide a suitable substrate for the excision machinery. In
398 support of the latter hypothesis, several chromatin remodelers and histone chaperones are known to
399 control PDE in *P. tetraurelia* (Ignarski et al. 2014; de Vanssay et al. 2020; Singh et al. 2022), but the
400 temporal and mechanistic details of their action remain to be precisely understood.

401 The existence of a sequential DNA elimination program may provide *Paramecium* with a
402 peculiar mechanism to fine-tune zygotic gene expression during MAC development. Developmental
403 regulation was previously proposed for genes located inside IESs or embedded in imprecisely
404 eliminated MIC-limited regions (Sellis et al. 2021): such germline-specific genes may be expressed

405 when zygotic transcription starts in the new MAC, until their encompassing DNA is removed from the
406 genome. PDE-mediated regulation of germline-limited genes was demonstrated in the ciliate *Euplotes*
407 *crassus* for a development-specific *de novo* telomerase gene (Karamysheva et al. 2003) and in
408 *Tetrahymena thermophila* for the gene encoding Tpb6p, a protein involved in excision of intragenic
409 IESs (Feng et al. 2017). It has also been reported in other organisms that eliminate germline-specific
410 genes during development (Smith et al. 2012; Wang et al. 2012; Chen et al. 2014; Torgasheva et al.
411 2019; Kinsella et al. 2019). In *Paramecium*, IESs could block zygotic gene expression as long as they
412 are present within coding sequences or in gene regulatory regions, as suggested for the *PTIW110* gene
413 (Furrer et al. 2017). An even more sophisticated regulatory scheme could be proposed with a first IES
414 excision event turning on an anlagen-specific gene that would subsequently be turned off by a second
415 DNA elimination event. In the future, monitoring the timing of PDE in other *Paramecium* species and
416 annotating the sequential versions of the rearranged genome should allow us to assess whether
417 temporal control of IES excision has been conserved during evolution and to what extent it may
418 contribute to gene regulation.

419

420 **DNA elimination timing reveals evolutionary optimization of TE-derived sequences for efficient** 421 **excision**

422 DNA and RNA transposons that have colonized the *Paramecium* germline genome during
423 evolution are eliminated in an imprecise manner from the new MAC during PDE (Arnaiz et al. 2012;
424 Guérin et al. 2017). We report here that imprecise elimination of TEs and other MIC-limited regions
425 occurs at a late stage during MAC development (DEV3 to DEV4). TE elimination was previously
426 shown to depend upon scnRNA-driven deposition of H3K9me3 and H3K27me3 histone marks, which
427 are enriched on TEs at a developmental stage corresponding to DEV2 and accumulate in the anlagen
428 up to DEV3 (Lhuillier-Akakpo et al. 2014; Frapporti et al. 2019). The deposition of heterochromatin
429 marks thus appears to be a late process, which might explain why TEs and other MIC-limited
430 sequences are eliminated at a late developmental stage.

431 The *Paramecium* MIC genome harbors TEs belonging to different families, most of which are
432 eliminated imprecisely during MAC development (Guérin et al. 2017). *Paramecium* IESs have been

433 proposed to originate from *Tc1/mariner* TEs, a particular family of transposons that duplicate their TA
434 target site upon integration into the germline genome (Arnaiz et al. 2012). Target site duplication
435 generates potential Pgm cleavage sites at the boundaries of newly inserted *Tc1/mariner* copies, which
436 has provided these TA-flanked TEs with the ability to be excised precisely and behave as IESs right
437 after their integration. We propose that, thanks to this ability, only *Tc1/mariner* TEs could be
438 maintained within genes in the germline and were allowed to decay, giving rise to extant IESs. This
439 evolutionary scenario was enriched by the finding that a handful of multicopy non-coding IESs were
440 recently mobilized in *trans* by transposases expressed from active TEs (Sellis et al. 2021). These
441 newly inserted IES copies are thought to evolve under the same constraints as all other IESs with
442 regard to their precise somatic excision.

443 In *P. tetraurelia*, IESs have shortened down to a minimal size range of 25 to 33-bp,
444 representing ~30% of all IESs. A phylogenetic analysis of IESs across *P. aurelia* species showed that
445 shortening has been accompanied by a switch in their excision mechanism. Indeed, the most recently
446 inserted IESs (*i.e.* the youngest) were shown to depend on scnRNAs and heterochromatin marks,
447 while old IESs have become independent of these epigenetic factors (Sellis et al. 2021). Here we show
448 that late excised IESs tend to be the youngest and that, similar to their TE ancestors, their elimination
449 depends on scnRNAs and histone marks. In addition, excision of late IESs tends to depend on the
450 presence of iesRNAs, which have been proposed to boost excision through a positive-feedback loop
451 (Sandoval et al. 2014). The stimulatory contribution of iesRNAs might explain why excision of late
452 IESs precedes imprecise elimination of TEs and other MIC-limited sequences during MAC
453 development. We also report that early excised IESs tend to be the oldest and are enriched for smaller
454 sizes (54.5% belong to the 25 to 33-bp peak). They are also mostly independent of sRNAs and
455 heterochromatin marks and tend to be the least error-prone. Our data therefore provide experimental
456 support to the proposed evolutionary scenario of *Paramecium* IESs, showing that their excision timing
457 reflects their evolutionary age (Fig. 5B).

458 We provide evidence that IESs have evolved through optimization for efficient excision,
459 combining an early and accurate excision process. Closer analysis of the intrinsic properties of very
460 early excised IESs furthermore revealed a strong nucleotide sequence signal at their ends, which varies

461 according to IES size (TATAG for 25 to 33-bp IESs, TACAG for 42 to 140-bp IESs). In contrast, late
462 excised IESs only exhibit a conserved TA dinucleotide at their ends. These observations suggest that
463 acquisition of a stronger sequence motif has allowed "optimized" IESs to loosen their requirement for
464 sRNAs for excision. Sequence-dependent determination of efficient excision would explain the
465 previous observation that 25 to 33-bp MAC genome segments flanked by terminal TATAG inverted
466 repeats are under-represented in the somatic MAC genome (Swart et al. 2014), possibly because such
467 sequences are highly excision-prone. MAC genome segments of any size flanked by terminal TACAG
468 inverted repeats are overall poorly represented as well in the *Paramecium* genome, thus precluding
469 their harmful excision (Swart et al. 2014). The present study therefore points to the joint contribution
470 of IES nucleotide sequence and size as intrinsic determinants of efficient IES excision. Several
471 hypotheses might explain how these determinants could work. By facilitating the formation of
472 particular DNA structures, they could help to target specific sequences for elimination, either through
473 a passive mechanism involving nucleosome exclusion to increase their accessibility, or by actively
474 promoting the assembly of the Pgm-endonuclease complex. The IES size-dependent consensus
475 sequences might also be related to a distinct spatial organization of IES ends within the excision
476 complex formed for very short vs long IESs (Arnaiz et al. 2012). Another non-exclusive hypothesis
477 might be that conserved sequence motifs help to position the Pgm catalytic domain on its cleavage
478 sites. Why two different sequence logos have evolved for different sizes of early excised IES remains
479 to be investigated. It could be linked to preferential recognition by different subunits of the excision
480 complex (e.g. PgmLs) which are all co-expressed with Pgm (Bischerour et al. 2018).

481 Studying the *Paramecium* model, with its nuclear dimorphism and ability to precisely excise
482 TE-related IESs even when inserted inside coding sequences, provides a unique opportunity to
483 monitor how TEs have degenerated within their host genomes. Further work on *Paramecium* PDE will
484 make it possible to decipher the evolutionary and mechanistic switch from sRNA- and
485 heterochromatin-mediated TE silencing to efficient elimination of TE-related sequences from the
486 genome. The characterization of a set of efficiently excised IESs, whose excision has become
487 independent of sRNAs and the heterochromatin pathway, paves the way to future biochemical studies

488 that will address the longstanding question of how domesticated PiggyBac transposases are recruited
489 to specific DNA cleavage sites to carry out precise DNA excision.

490

491

492 **Methods**

493 **Cell growth and autogamy time-courses**

494 Culture of *P. tetraurelia* wild type 51 new (Gratias and Bétermier 2003) or its mutant derivative 51
495 *nd7-1* (Dubois et al. 2017) was performed using standard conditions. Briefly, cells were grown in
496 medium made of wheat grass infusion (WGP) inoculated with *Klebsiella pneumoniae* and
497 supplemented with β -sitosterol (0.8 μ g/ml) prior to use (Beisson et al. 2010). For autogamy time-
498 courses, cells (~20-30 vegetative fissions) were seeded at a final concentration of 250 cells/mL in
499 inoculated WGP medium with an OD_{600nm} adjusted to 0.1. Autogamy was triggered by starvation the
500 next day. We performed 6 independent autogamy time-courses (tc1 to tc6). For each time-course, the
501 T0 time-point was defined as the time (in hours) when 50% of the cells in the population have a
502 fragmented MAC. We further defined 5 developmental stages: DEV1 (T2.5-T3), DEV2 (T7-T12),
503 DEV3 (T20-T24), DEV4 (T30), DEV5 (T48), with time-points following T0 as previously described
504 (Armaiz et al. 2017). At each selected time-point, 0.7-2 L of culture (at a concentration of 1500-3500
505 cells/mL) were processed for nuclear preparation or 30 mL for whole cell immunofluorescence. To
506 inhibit DNA replication during autogamy, aphidicolin (Sigma-Aldrich, ref. A0781) was added at T2.5
507 at a final concentration of 15 μ M, and the same volume was added a second time at T10. Cells were
508 harvested and nuclei were isolated at T20. For all time-courses, the survival of post-autogamous
509 progeny was tested as described before (Dubois et al. 2017).

510

511 **Immunofluorescence analysis**

512 A peptide corresponding to PgmL1 amino acid sequence 1 to 266 and carrying a C-terminal His tag
513 was used for guinea pig immunization (Proteogenix). Sera were purified by antigen affinity
514 purification to obtain highly specific α -PgmL1-GP antibodies (0.8 mg/mL). RNAi targeting the
515 *PGML1* gene during autogamy, immunofluorescence labeling of whole cells and quantification of
516 PgmL1 signal were performed as described previously (Bischerour et al. 2018). Cells were extracted
517 with ice-cold PHEM (60 mM PIPES, 25 mM HEPES, 10 mM EGTA, 2 mM MgCl₂ pH 6.9) + 1%
518 Triton prior to fixation and immunostaining with α -PgmL1-GP (1:2000).

519

520 **Isolation of nuclei and immunostaining**

521 Nuclear preparations enriched in developing MACs were obtained as previously described (Arnaiz et
522 al. 2012) with few modifications: the cell pellet was resuspended in 6-10 volumes of lysis buffer (0.25
523 M sucrose, 10 mM MgCl₂, 10 mM Tris pH 6.8, 0.2% Nonidet P-40) supplemented with 2x Protease
524 Inhibitor Cocktail Set I (PICS, Cabiochem, ref. 539131), kept on ice for 15 min and disrupted with a
525 Potter-Elvehjem homogenizer (100 to 400 strokes). Lysis efficiency was monitored with a Zeiss
526 Lumar.V12 fluorescence stereo-microscope, following addition of 66 µg/mL DAPI. Nuclei were
527 collected through centrifugation at 1000 g for 2 min and washed four times with 10 volumes of
528 washing buffer (0.25 M sucrose, 10 mM MgCl₂, 10 mM Tris pH 7.4). The nuclear pellet was either
529 diluted 2-fold in washing buffer containing glycerol (13% final concentration) and frozen as aliquots
530 at -80°C or diluted 2-fold in washing buffer supplemented with 2x PICS and loaded on top of a 3-mL
531 sucrose (2.1 M) layer before ultra-centrifugation in a swinging rotor for 1 h at 210,000 g. After gentle
532 washes, the pellet was resuspended in 1 volume of washing buffer containing glycerol (13% final
533 concentration) and frozen as aliquots at -80°C. New MAC labeling was adapted from a published
534 method (Sardo et al. 2017). Nuclear preparations were immunostained on ice for 1 h in TBS (10 mM
535 Tris pH 7.4, 0.15 M NaCl) + 3% BSA containing α-PgmL1-GP (1:1000). Nuclei were washed twice
536 in TBS + 3% BSA and stained for 45 min with Alexa Fluor (AF) 488-conjugated goat anti-guinea pig
537 IgG (1:500, ThermoFisher Scientific). Nuclei were finally washed twice in TBS + 3% BSA and
538 resuspended in PI/RNase staining buffer (BD Pharmingen, ref. 550825). All centrifugation steps were
539 performed at 500 g for 1 min at 4°C. Samples were kept in the dark at 4°C until processing.

540

541 **Flow cytometry**

542 Stained nuclei were filtered through sterile 30 µm cell strainers (Sysmex filters, CellTrics® ref. 04-
543 004-2326) and processed for flow cytometry. Immunostained nuclei were analyzed on a CytoFlex S
544 cytometer (Beckman Coulter) with a 488 nm laser for scatter measurements (Forward scatter, or FSC,
545 and Side scatter, or SSC) and AF488 excitation, and a 561 nm laser for PI excitation. AF488 and PI
546 staining signals were respectively collected using a 525/40 nm band pass filter and a 610/20 nm band

547 pass filter. Immunostained nuclei were sorted on a Moflow Astrios EQ cell sorter (Beckman Coulter)
548 with a 488-nm laser for scatter measurements (Forward Scatter, or FCS, and Side Scatter, or SSC) and
549 AF488 excitation, and a 561 nm laser for PI excitation. AF488 and PI staining signals were
550 respectively collected using a 526/52 nm band pass filter and a 614/20 nm band pass filter. Phosphate
551 Buffered Saline-like (Puraflo Sheath Fluid, Beckman Coulter) was used as sheath and run at a
552 constant pressure of 10 or 25 PSI. Frequency of drop formation was 26 or 43 kHz. Purify mode was
553 used for sorting in order to reach a maximum rate of purity (>95%). The instrument used a 100 µm
554 nozzle. A threshold on the PI signal was optimized to increase collecting speed (~1000 events per
555 second). Data were collected using Summit software (Beckman Coulter). Nuclei were first gated based
556 on their Side Scatter (SSC-area) and high PI signal (PI-area), and sorted according to their AF488
557 signal. AF488-positive events were backgated onto SSC vs PI to optimize the gating. Doublets were
558 discarded using PI-area and PI-height signals. Nuclei (<30,000) were collected into 100 µl of Buffer
559 AL (QIAamp DNA Micro Kit, QIAGEN) and immediately lysed by pulse-vortexing. Final volume
560 was adjusted to 200 µL with PI/RNase staining buffer. We confirmed that the FANS procedure yields
561 pure anlagen in an experiment (tc3) in which old MAC fragments contained a marker transgene absent
562 from the anlagen (Supplemental Fig. S3A-C and Supplemental Methods).

563

564 **Estimation of new MAC DNA content by flow cytometry**

565 Estimation of the absolute DNA content (C-value, in Mbp) in the new MAC populations was based on
566 a previously described method (Bourge et al. 2018). The DNA content was calculated using the linear
567 relationship between the fluorescent signal from the new MAC peaks and a known internal standard
568 (tomato nuclei, *Solanum lycopersicum* L. cv. Montfavet 63-5, 2C=1946 Mbp). Briefly, leaves were
569 chopped with a razor blade in a Petri dish with PI/RNase staining buffer, filtered through 30-µm cell
570 strainers and added in a constant ratio to an aliquot of stained and filtered *Paramecium* nuclei. The C-
571 value ($C_{\text{new MACs}}$) for each new MAC subpopulation was calculated using its PI Mean Fluorescence
572 Intensity ($\text{MFI}_{\text{new MACs}}$), the PI Mean Fluorescence Intensity of the 2C tomato standard ($\text{MFI}_{\text{standard}}$) and
573 the 2C-value of the tomato standard ($2C_{\text{standard}}$) (Supplemental Figure S2):

$$574 \quad C_{\text{new MACs}} = \text{MFI}_{\text{new MACs}} \times 2C_{\text{standard}} / \text{MFI}_{\text{standard}}$$

575 The endoreplication level for each new MAC population (C-level) was further estimated by dividing
576 the C-value for each new MAC population by the DNA content of the unarranged *P. tetraurelia*
577 MIC genome (1C=100 Mbp) (Guérin et al. 2017; Sellis et al. 2021): $C\text{-level} = C_{\text{new MACs}} / 1C_{\text{mic}}$

578

579 **Genomic DNA extraction and high-throughput sequencing**

580 DNA extraction was performed using the QIAamp DNA Micro Kit (Qiagen) as recommended by the
581 manufacturer, with minor modifications. Following a 10-min incubation with proteinase K (2 mg/mL)
582 in buffer AL, the nuclear lysate was directly loaded onto the purification column. Elution was
583 performed with 20-50 μ L Buffer AE in DNA LoBind Eppendorf Tubes. DNA concentration was
584 determined using the QBit High Sensitivity kit (Invitrogen) before storage at -20°C . Sequencing
585 libraries were prepared using 1.5 to 8.5 ng of DNA with the TruSeq NGS Library Prep kit from
586 Westburg (WB9024) following manufacturer's instructions. Alternatively, genomic DNA was
587 fragmented with the S220 Focused-ultrasonicator (Covaris). Fragments were processed with NEBNext
588 Ultra II End Prep Reagents (NEB #E7546) and TruSeq adapters were ligated using the NEBNext
589 Quick Ligation kit (NEB #E6056). Libraries were amplified by PCR using Kapa HiFi DNA
590 polymerase (10-14 cycles). Library quality was checked with an Agilent Bioanalyzer instrument
591 (Agilent High Sensitivity DNA kit). Sequencing was performed on 75-75bp paired-end runs, with an
592 Illumina NextSeq500/550 instrument, using the NextSeq 500/550 MID output cycle kit.
593 Demultiplexing was performed with bcl2fastq2-2.18.12 (<https://emea.support.illumina.com/>) and
594 adapters were removed with Cutadapt 1.15 (Martin 2011); only reads longer than 10 bp were retained.

595

596 **Software and R packages**

597 Sequencing reads were mapped on genome references using Bowtie 2 (v2.2.9 --local --X 500)
598 (Langmead and Salzberg 2012). The resulting alignments were analyzed using SAMtools (v1.9) (Li et
599 al. 2009), ParTIES (v1.05 <https://github.com/oarnaiz/ParTIES>) (Denby Wilkes et al. 2016) and
600 BEDtools (v2.26) (Quinlan and Hall 2010). R (v4) packages were used to generate images (ggplot2
601 v3.3.5; ComplexHeatmap v2.6.2; GenomicRanges v1.42) (R Core Team 2021; Wickham 2016; Gu et

602 al. 2016; Lawrence et al. 2013). The IES sequence end logos were generated using WebLogo (v3.6.0 -
603 -composition 0.28 --units bits) (Crooks et al. 2004).

604

605 **Reference genomes and datasets**

606 Paired-end sequencing data were mapped on *P. tetraurelia* strain 51 MAC (ptetraurelia_mac_51.fa),
607 MAC+IES (ptetraurelia_mac_51_with_ies.fa) or MIC (ptetraurelia_mic2.fa) reference genomes
608 (Arnaiz et al. 2012; Guérin et al. 2017). Gene annotation v2.0
609 (ptetraurelia_mac_51_annotation_v2.0.gff3), IES annotation v1
610 (internal_eliminated_sequence_PGM_ParTIES.pt_51.gff3) and TE annotation v1.0
611 (ptetraurelia_mic2_TE_annotation_v1.0.gff3) were used in this study (Arnaiz et al. 2012, 2017;
612 Guérin et al. 2017). All files are available from the ParameciumDB download section
613 (<https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/tetraurelia/51/>) (Arnaiz et al. 2020).
614 DNA sequencing data of *Paramecium* cells depleted of Ez11, TFIS4, Dcl2/3 or Dcl5 were previously
615 published (Lhuillier-Akakpo et al. 2014; Maliszewska-Olejniczak et al. 2015; Sandoval et al. 2014).
616 ParTIES (MIRET module) was used to determine the IESs that were significantly retained compared
617 to the control (Denby Wilkes et al. 2016). An IES is considered to be dependent on the depleted factor
618 for its excision (31,505; 20,524; 3,439 and 2,475 IESs sensitive to *EZL1*, *TFIS4*, *DCL2/3* and *DCL5*
619 RNAi, respectively), if at least one IES boundary in at least one replicate shows significant retention.

620

621 **Excision score calculation and IES classification**

622 Mapping of sequencing reads on the MAC and the MAC+IES references was used to calculate an IES
623 Excision Score ($ES = IES^- / (IES^+ + IES^-)$) using ParTIES (MIRET default parameters). An ES of 0
624 means no excision and an ES of 1 means complete IES excision. The violin plots show the distribution
625 of the mean ES score for the two IES boundaries of all IESs. Excision profile classification was
626 carried out on the 44928 annotated IESs, after removing 543 IESs with $ES < 0.8$, in at least one FRAG
627 sample, which indicates imperfect excision in the old MAC (group defined as « None »). *K*-means
628 clustering (iter.max=100, *k*-means R function from "stats" package) was used to define 4 groups based
629 on the ES in all conditions.

630

631 **TE and genome coverage**

632 The mean sequencing depth (SAMtools depth -q 30 -Q 30), normalized by the number of reads
633 mapped on the MIC reference genome, was calculated on TE copies (500 nt min length and localized
634 on MIC contigs > 2kb) and IESs. Only fully mapped reads overlapping at least 4 nucleotides of the
635 annotated feature were considered. As previously described (Guérin et al. 2017), the same window
636 coverage approach was used to estimate genome coverage at each time point. The coverage (multicov
637 -q 30) was calculated for non-overlapping 1-kb windows, then normalized by the total number of
638 mapped reads (RPM). An empirical cutoff of 2.5 RPM was used to decide if the window is covered or
639 not.

640

641 **Detection of *de novo* telomere addition sites**

642 *De novo* telomere addition sites were identified on the MIC genome, with the requirement of at least 3
643 consecutive repeats of either G₄T₂ or G₃T₃ on mapped reads. A telomere addition site was identified if
644 the read alignment stops at the exact position where the telomeric repeat starts. The number of
645 telomere addition sites was normalized by the number of reads mapped on the MIC genome.

646

647 **IES-IES junctions**

648 The ParTIES Concatemer module, developed for this study, was used with default parameters to
649 identify concatemers of excised IESs. Reads were recursively mapped to the IES sequences, as shown
650 in Supplemental Fig. S5A. At each round, reads are mapped to IES sequences and selected if the
651 alignment begins or ends at an IES extremity. If the read is partially aligned, then the unmapped part
652 of the read is re-injected into the mapping and the selection procedure continues until the entire read
653 has been mapped.

654

655 **IES excision errors**

656 The ParTIES MILORD module was used with the MAC+IES reference genome to identify IES
657 excision errors. Only error types described in Supplemental Fig. S7A were considered. The number of

658 non-redundant errors was normalized by the number of mapped reads. PCR duplicates were removed
659 using SAMtools rmdup.

660

661 **Data access**

662 Genes used in this study are accessible in ParameciumDB as follows: *PGM* (PTET.51.1.G0490162),
663 *PGML1* (PTET.51.1.G0110267), *EZLI* (PTET.51.1.G1740049), *TFIIS4* (PTET.51.1.G0900102),
664 *DCL2* (PTET.51.1.G0210241), *DCL3* (PTET.51.1.G0990073), *DCL5* (PTET.51.1.G0070121), *ND7*
665 (PTET.51.1.G0050374) (Arnaiz et al. 2020).

666 The sequencing data generated for this study have been submitted to the ENA database
667 (<https://www.ebi.ac.uk/ena/browser/home>) under accession number PRJEB49315.

668 The statistical data, scripts (Supplemental Codes) and raw images have been deposited at Zenodo
669 (<https://doi.org/10.5281/zenodo.6534539>).

670 The cytometry data generated in this study have been submitted to the FlowRepository database
671 ([http://flowrepository.org/id/RvFrl4FUJTnaAIDEsEqK3MzxKwQZpkfp7yqzGGsco3tuuLfuAHKrPI2](http://flowrepository.org/id/RvFrl4FUJTnaAIDEsEqK3MzxKwQZpkfp7yqzGGsco3tuuLfuAHKrPI2fP65KehpH)
672 *fP65KehpH*).

673

674 **Competing interest statement**

675 The authors declare no competing interest.

676

677 **Acknowledgments**

678 We would like to thank Cindy Mathon and Pascaline Tirand for their technical assistance and all
679 members of the Bétermier laboratory for stimulating and fruitful discussions. Special thanks to Joël
680 Acker, Julien Bischerour and Linda Sperling for critical reading of the manuscript. This study was
681 supported by intramural funding from the Centre National de la Recherche (CNRS) and by grants
682 from the *Agence Nationale de la Recherche* (LaMarque ANR-18-CE12-0005-02 & CURE ANR-21-
683 CE12-0019-01 to M. Bétermier, POLYCHROME ANR-19-CE12-0015 to O.A.) and the *Fondation*
684 *pour la Recherche Médicale* (FRM EQU202103012766 to M. Bétermier). The present work has
685 benefited from the expertise of the Imagerie-Gif core facility, supported by the *Agence Nationale de la*

686 *Recherche* (ANR-11-EQPX-0029/Morphoscope, ANR-10-INBS-04/FranceBioImaging,
687 ANR-11-IDEX-0003-02/ Saclay Plant Sciences). We acknowledge the sequencing and bioinformatics
688 expertise of the I2BC High-throughput sequencing facility, supported by *France Génomique* (funded
689 by the French National Program “Investissement d’Avenir” ANR-10-INBS-09).

690

691 **Author contributions**

692 C.Z., M.Bourge, M.Bétermier, O.A, V.R.: designed research, analyzed data.

693 C.Z, K.G., L.E., M.Bourge, N.M., O.A, V.R., Y.J.: performed research.

694 O.A.: developped bioinformatic pipelines and conducted bioinformatic data analysis.

695 C.Z., K.G., M.Bourge, M.Bétermier, O.A.,V.R.: wrote the paper.

696

697 **References**

- 698 Abello A, Régnier V, Arnaiz O, Le Bars R, Betermier M, Bischerour J. 2020. Functional
699 diversification of *Paramecium* Ku80 paralogs safeguards genome integrity during precise
700 programmed DNA elimination. *PLoS Genet* **16**: e1008723.
- 701 Allen SE, Hug I, Pabian S, Rzeszutek I, Hoehener C, Nowacki M. 2017. Circular concatemers of
702 ultra-short DNA segments produce regulatory RNAs. *Cell* **168**: 990–999.
- 703 Allen SE, Nowacki M. 2020. Roles of Noncoding RNAs in Ciliate Genome Architecture. *J Mol Biol.*
704 <http://www.ncbi.nlm.nih.gov/pubmed/31926952>.
- 705 Arnaiz O, Mathy N, Baudry C, Malinsky S, Aury JM, Denby Wilkes C, Garnier O, Labadie K,
706 Lauderdale BE, Le Mouel A, et al. 2012. The *Paramecium* germline genome provides a niche
707 for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. *PLoS*
708 *Genet* **8**: e1002984.
- 709 Arnaiz O, Meyer E, Sperling L. 2020. ParameciumDB 2019: integrating genomic data across the
710 genus for functional and evolutionary biology. *Nucleic Acids Res* **48**: D599–D605.
- 711 Arnaiz O, Van Dijk E, Bétermier M, Lhuillier-Akakpo M, de Vanssay A, Duharcourt S, Sallet E,
712 Gouzy J, Sperling L. 2017. Improved methods and resources for paramecium genomics:
713 transcription units, gene annotation and gene expression. *BMC Genomics* **18**: 483.
- 714 Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V,
715 Aiach N, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate
716 *Paramecium tetraurelia*. *Nature* **444**: 171–8.
- 717 Baroin A, Prat A, Caron F. 1987. Telomeric site position heterogeneity in macronuclear DNA of
718 *Paramecium primaurelia*. *Nucleic Acids Res* **15**: 1717–28.
- 719 Bates AD, Noy A, Piperakis MM, Harris SA, Maxwell A. 2013. Small DNA circles as probes of
720 DNA topology. *Biochem Soc Trans* **41**: 565–570.
- 721 Baudry C, Malinsky S, Restituto M, Kapusta A, Rosa S, Meyer E, Bétermier M. 2009. PiggyMac, a
722 domesticated piggyBac transposase involved in programmed genome rearrangements in the
723 ciliate *Paramecium tetraurelia*. *Genes Dev* **23**: 2478–2483.
- 724 Beisson J, Bétermier M, Bré M-H, Cohen J, Duharcourt S, Duret L, Kung C, Malinsky S, Meyer E,
725 Preer JR, et al. 2010. *Paramecium tetraurelia*: the renaissance of an early unicellular model.
726 *Cold Spring Harb Protoc* **2010**: pdb.emo140.
- 727 Bennetzen JL, Park M. 2018. Distinguishing friends, foes, and freeloaders in giant genomes. *Curr*
728 *Opin Genet Dev* **49**: 49–55.
- 729 Berger JD. 1986. Autogamy in *Paramecium*. Cell cycle stage-specific commitment to meiosis. *Exp*
730 *Cell Res* **166**: 475–485.
- 731 Berger JD. 1973. Nuclear differentiation and nucleic acid synthesis in well-fed exconjugants of
732 *Paramecium aurelia*. *Chromosoma* **42**: 247–268.
- 733 Bétermier M, Bertrand P, Lopez BS. 2014. Is non-homologous end-joining really an inherently
734 error-prone process? *PLoS Genet* **10**: e1004086.
- 735 Bétermier M, Borde V, Villartay J-P de. 2020. Coupling DNA Damage and Repair: an Essential
736 Safeguard during Programmed DNA Double-Strand Breaks? *Trends in Cell Biology* **30**: 87–96.
- 737 Bétermier M, Duharcourt S. 2014. Programmed rearrangement in ciliates: *Paramecium*. *Microbiol*
738 *Spectr* **2**: MDNA3-0035-2014.

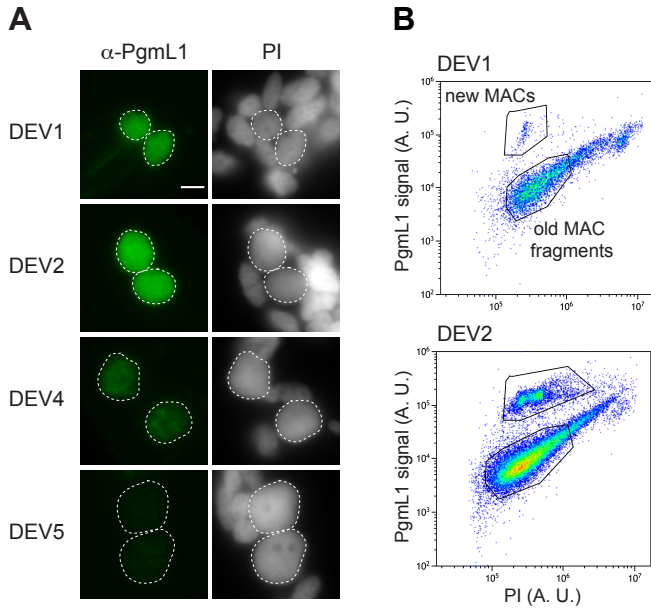
- 739 B  termier M, Duharcourt S, Seitz H, Meyer E. 2000. Timing of Developmentally Programmed
740 Excision and Circularization of Paramecium Internal Eliminated Sequences. *Mol Cell Biol* **20**:
741 1553–1561.
- 742 Bischerour J, Bhullar S, Denby Wilkes C, Regnier V, Mathy N, Dubois E, Singh A, Swart E, Arnaiz
743 O, Sperling L, et al. 2018. Six domesticated PiggyBac transposases together carry out
744 programmed DNA elimination in Paramecium. *Elife* **7**: e37927.
- 745 Bleykasten-Grosshans C, Neueglise C. 2011. Transposable elements in yeasts. *C R Biol* **334**: 679–
746 86.
- 747 Bourge M, Brown SC, Siljak-Yakovlev S. 2018. Flow cytometry as tool in plant sciences, with
748 emphasis on genome size and ploidy level assessment. *GenApp* **2**: 1.
- 749 Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007. Discrete
750 small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**:
751 1089–103.
- 752 Byrnes JJ. 1984. Structural and functional properties of DNA polymerase delta from rabbit bone
753 marrow. *Mol Cell Biochem* **62**: 13–24.
- 754 Capy P. 2021. Taming, Domestication and Exaptation: Trajectories of Transposable Elements in
755 Genomes. *Cells* **10**. <http://www.ncbi.nlm.nih.gov/pubmed/34944100>.
- 756 Chalker DL, Meyer E, Mochizuki K. 2013. Epigenetics of ciliates. *Cold Spring Harb Perspect Biol*
757 **5**: a017764.
- 758 Chen X, Bracht JR, Goldman AD, Dolzhenko E, Clay DM, Swart EC, Perlman DH, Doak TG, Stuart
759 A, Amemiya CT, et al. 2014. The Architecture of a Scrambled Genome Reveals Massive Levels
760 of Genomic Rearrangement during Development. *Cell* **158**: 1187–1198.
- 761 Cheng CH, Kuchta RD. 1993. DNA polymerase epsilon: aphidicolin inhibition and the relationship
762 between polymerase and exonuclease activity. *Biochemistry* **32**: 8568–8574.
- 763 Cheng CY, Orias E, Leu JY, Turkewitz AP. 2020. The evolution of germ-soma nuclear
764 differentiation in eukaryotic unicells. *Curr Biol* **30**: R502–R510.
- 765 Choi JY, Lee YCG. 2020. Double-edged sword: The evolutionary consequences of the epigenetic
766 silencing of transposable elements. *PLoS Genet* **16**: e1008872.
- 767 Cosby RL, Chang NC, Feschotte C. 2019. Host-transposon interactions: conflict, cooperation, and
768 cooption. *Genes Dev* **33**: 1098–1116.
- 769 Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: A Sequence Logo Generator.
770 *Genome Res* **14**: 1188–1190.
- 771 de Vanssay A, Touzeau A, Arnaiz O, Frapporti A, Phipps J, Duharcourt S. 2020. The Paramecium
772 histone chaperone Spt16-1 is required for Pgm endonuclease function in programmed genome
773 rearrangements. *PLoS Genet* **16**: e1008949.
- 774 D  l  ris A, Berger F, Duharcourt S. 2021. Role of Polycomb in the control of transposable elements.
775 *Trends Genet* **37**: 882–889.
- 776 Denby Wilkes C, Arnaiz O, Sperling L. 2016. ParTIES: a toolbox for Paramecium interspersed DNA
777 elimination studies. *Bioinformatics* **32**: 599–601.
- 778 Deniz O, Frost JM, Branco MR. 2019. Regulation of transposable elements by DNA modifications.
779 *Nat Rev Genet* **20**: 417–431.

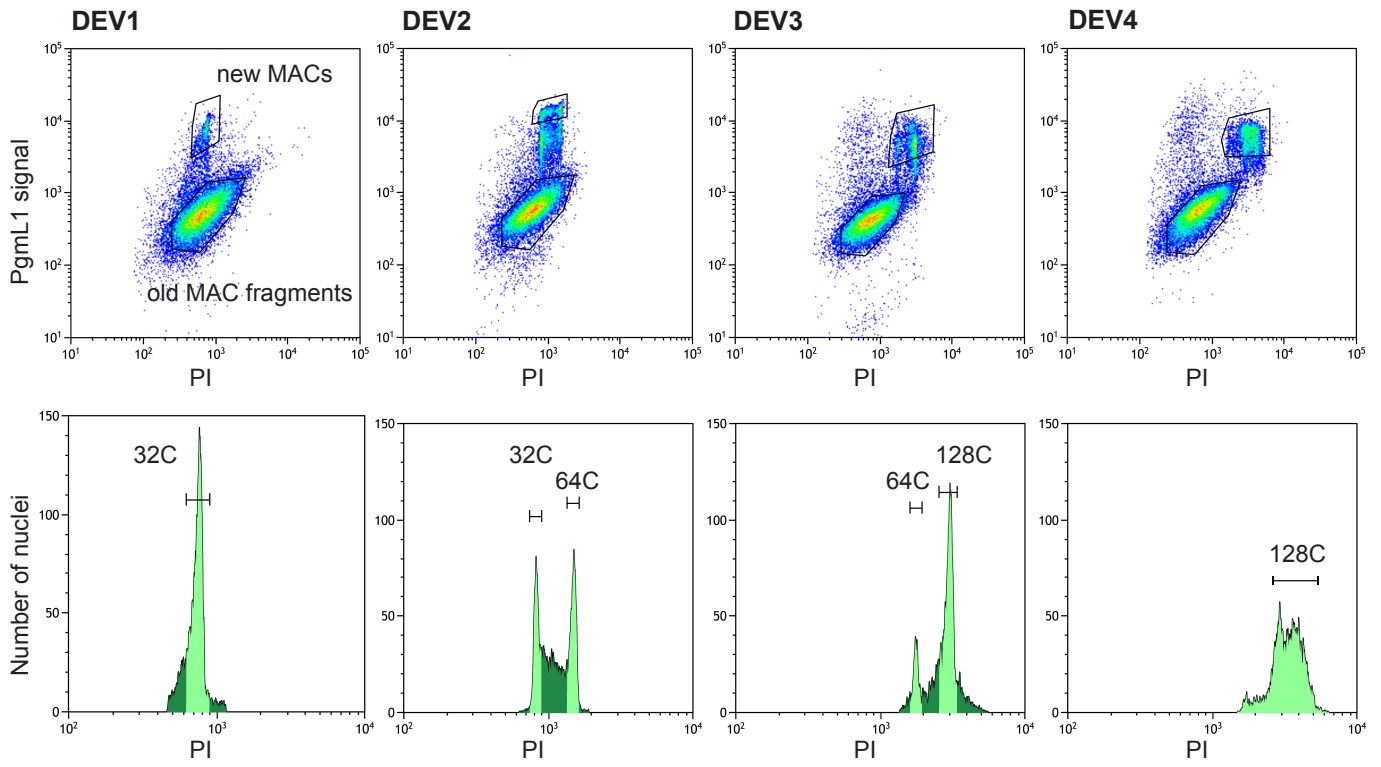
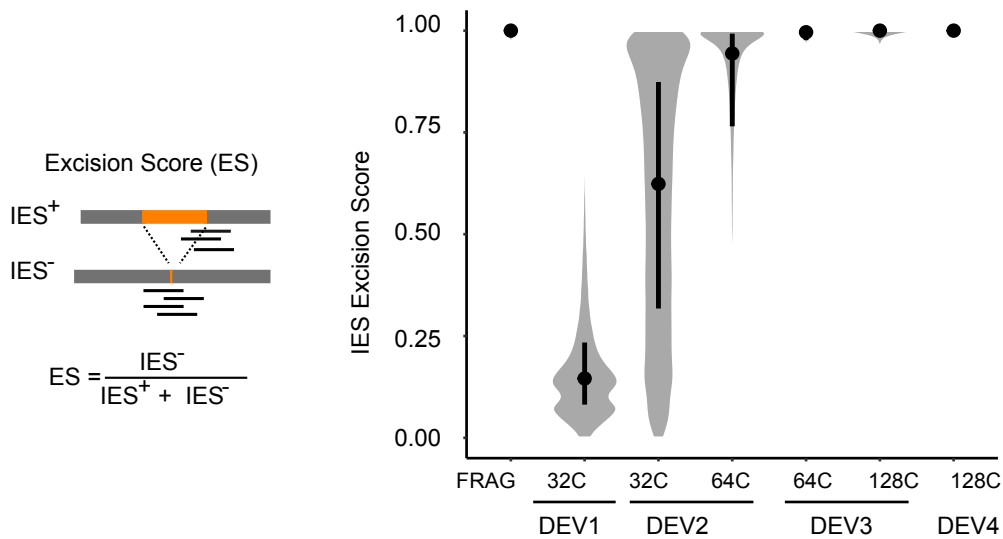
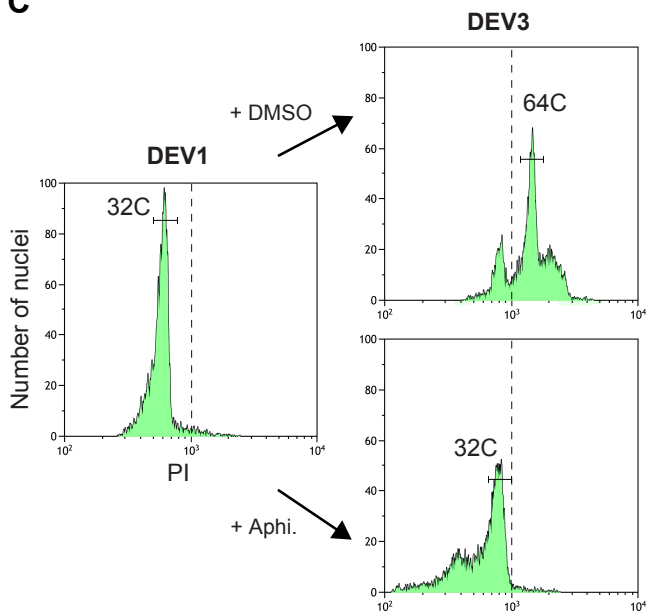
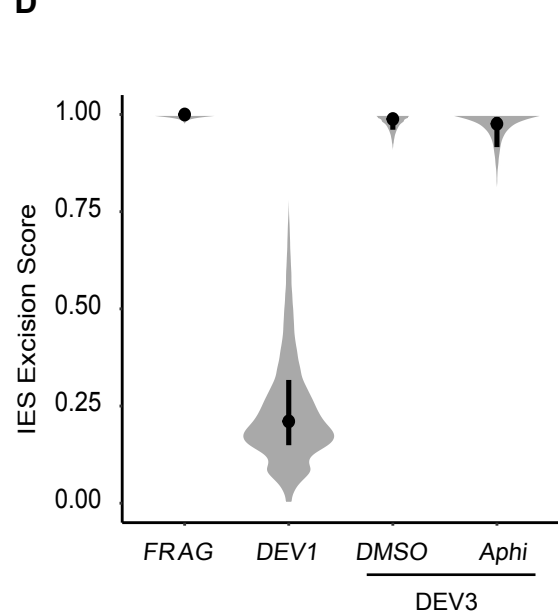
- 780 Dubois E, Bischerour J, Marmignon A, Mathy N, Regnier V, Betermier M. 2012. Transposon
781 invasion of the Paramecium germline genome countered by a domesticated PiggyBac
782 transposase and the NHEJ pathway. *Int J Evol Biol* **2012**: 436196.
- 783 Dubois E, Mathy N, Regnier V, Bischerour J, Baudry C, Trouslard R, Betermier M. 2017.
784 Multimerization properties of PiggyMac, a domesticated piggyBac transposase involved in
785 programmed genome rearrangements. *Nucleic Acids Res* **45**: 3204–3216.
- 786 Duharcourt S, Keller A-M, Meyer E. 1998. Homology-Dependent Maternal Inhibition of
787 Developmental Excision of Internal Eliminated Sequences in Paramecium tetraurelia. *Mol Cell*
788 *Biol* **18**: 7075–7085.
- 789 Duret L, Cohen J, Jubin C, Dessen P, Goût J-F, Mousset S, Aury J-M, Jaillon O, Noël B, Arnaiz O,
790 et al. 2008. Analysis of sequence variability in the macronuclear DNA of Paramecium
791 tetraurelia: a somatic view of the germline. *Genome Res* **18**: 585–596.
- 792 Feng L, Wang G, Hamilton EP, Xiong J, Yan G, Chen K, Chen X, Dui W, Plemens A, Khadr L, et
793 al. 2017. A germline-limited piggyBac transposase gene is required for precise excision in
794 Tetrahymena genome rearrangement. *Nucleic Acids Research* **45**: 9481–9502.
- 795 Frapporti A, Miro Pina C, Arnaiz O, Holoch D, Kawaguchi T, Humbert A, Eleftheriou E, Lombard
796 B, Loew D, Sperling L, et al. 2019. The Polycomb protein Ezh1 mediates H3K9 and H3K27
797 methylation to repress transposable elements in Paramecium. *Nat Commun* **10**: 2710.
- 798 Furrer DI, Swart EC, Kraft MF, Sandoval PY, Nowacki M. 2017. Two Sets of Piwi Proteins Are
799 Involved in Distinct sRNA Pathways Leading to Elimination of Germline-Specific DNA. *Cell*
800 *Rep* **20**: 505–520.
- 801 Görtz HD. 1988. *Paramecium*. Springer-Verlag, Berlin Heidelberg New York.
- 802 Gratias A, Bétermier M. 2001. Developmentally programmed excision of internal DNA sequences in
803 Paramecium aurelia. *Biochimie* **83**: 1009–1022.
- 804 Gratias A, Bétermier M. 2003. Processing of double-strand breaks is involved in the precise excision
805 of Paramecium IESs. *Mol Cell Biol* **23**: 7152–7162.
- 806 Gratias A, Lepere G, Garnier O, Rosa S, Duharcourt S, Malinsky S, Meyer E, Betermier M. 2008.
807 Developmentally programmed DNA splicing in Paramecium reveals short-distance crosstalk
808 between DNA cleavage sites. *Nucleic Acids Res* **36**: 3244–3251.
- 809 Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in
810 multidimensional genomic data. *Bioinformatics* **32**: 2847–2849.
- 811 Guérin F, Arnaiz O, Boggetto N, Denby Wilkes C, Meyer E, Sperling L, Duharcourt S. 2017. Flow
812 cytometry sorting of nuclei enables the first global characterization of Paramecium germline
813 DNA and transposable elements. *BMC Genomics* **18**: 327.
- 814 Hafez M, Hausner G. 2015. Convergent evolution of twintron-like configurations: One is never
815 enough. *RNA Biology* **12**: 1275–1288.
- 816 Hamilton EP, Kapusta A, Huvos PE, Bidwell SL, Zafar N, Tang H, Hadjithomas M, Krishnakumar
817 V, Badger JH, Caler EV, et al. 2016. Structure of the germline genome of Tetrahymena
818 thermophila and relationship to the massively rearranged somatic genome. *Elife* **5**.
819 <http://www.ncbi.nlm.nih.gov/pubmed/27892853>.
- 820 Ignarski M, Singh A, Swart EC, Arambasic M, Sandoval PY, Nowacki M. 2014. Paramecium
821 tetraurelia chromatin assembly factor-1-like protein PtCAF-1 is involved in RNA-mediated
822 control of DNA elimination. *Nucleic Acids Res* **42**: 11952–11964.

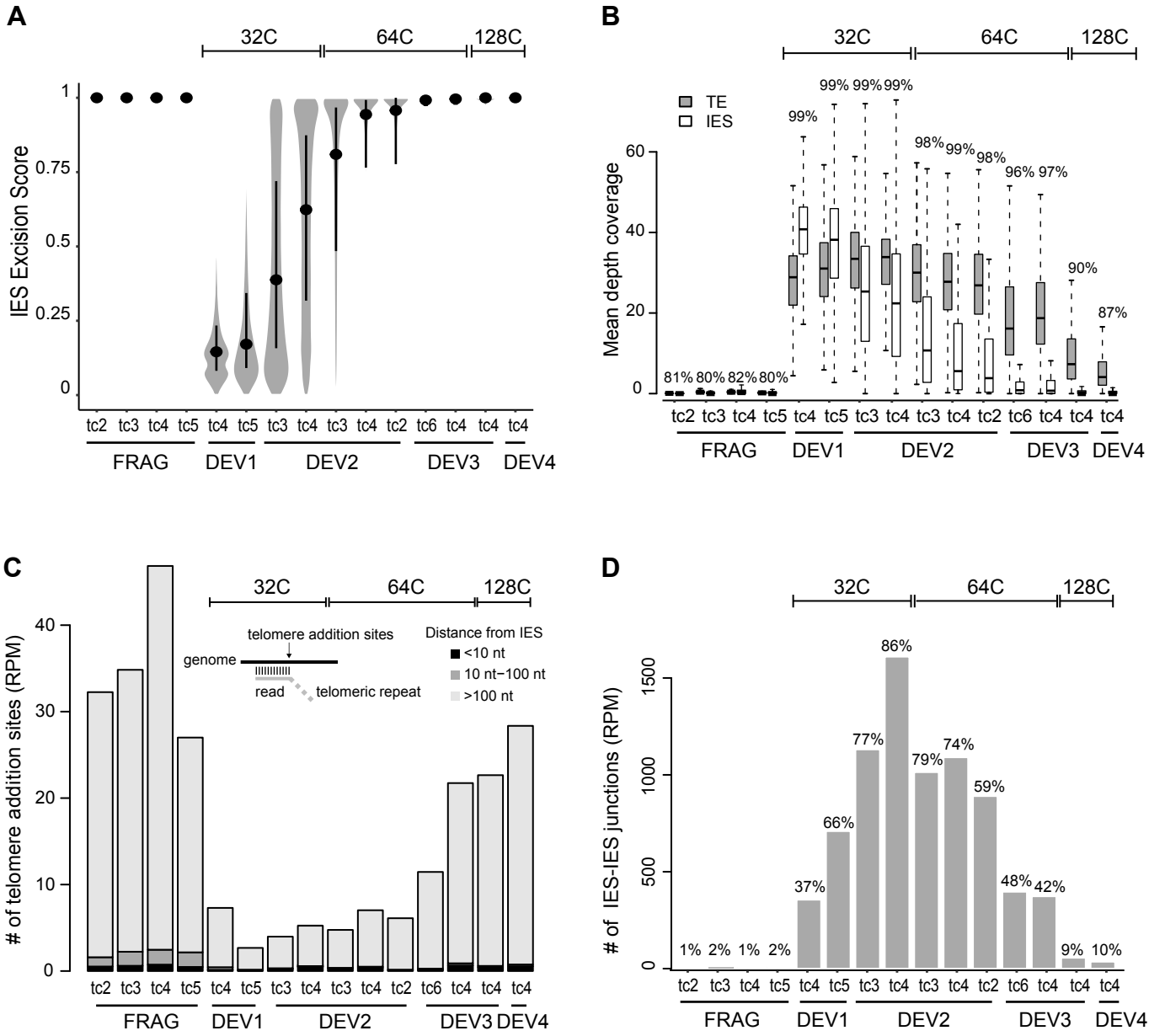
- 823 Kapusta A, Matsuda A, Marmignon A, Ku M, Silve A, Meyer E, Forney JD, Malinsky S, Bétermier
824 M. 2011. Highly Precise and Developmentally Programmed Genome Assembly in Paramecium
825 Requires Ligase IV-Dependent End Joining. *PLoS Genet* **7**.
826 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3077386/> (Accessed April 10, 2020).
- 827 Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals.
828 *Proc Natl Acad Sci U S A* **114**: E1460–E1469.
- 829 Karamysheva Z, Wang L, Shrode T, Bednenko J, Hurley LA, Shippen DE. 2003. Developmentally
830 Programmed Gene Elimination in *Euplotes crassus* Facilitates a Switch in the Telomerase
831 Catalytic Subunit. *Cell* **113**: 565–576.
- 832 Ketting RF, Haverkamp TH, van Luenen HG, Plasterk RH. 1999. Mut-7 of *C. elegans*, required for
833 transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and
834 RNaseD. *Cell* **99**: 133–41.
- 835 Kinsella CM, Ruiz-Ruano FJ, Dion-Côté A-M, Charles AJ, Gossmann TI, Cabrero J, Kappei D,
836 Hemmings N, Simons MJP, Camacho JPM, et al. 2019. Programmed DNA elimination of
837 germline development genes in songbirds. *Nat Commun* **10**: 5468.
- 838 Klobutcher LA, Herrick G. 1997. Developmental genome reorganization in ciliated protozoa: the
839 transposon link. *Progr Nucleic Acid Res Mol Biol* **56**: 1–62.
- 840 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–
841 359.
- 842 Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ.
843 2013. Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology*
844 **9**: e1003118.
- 845 Le Mouél A, Butler A, Caron F, Meyer E. 2003. Developmentally regulated chromosome
846 fragmentation linked to imprecise elimination of repeated sequences in paramecia. *Eukaryot*
847 *Cell* **2**: 1076–1090.
- 848 Lepère G, Bétermier M, Meyer E, Duharcourt S. 2008. Maternal noncoding transcripts antagonize
849 the targeting of DNA elimination by scanRNAs in *Paramecium tetraurelia*. *Genes Dev* **22**:
850 1501-1512.
- 851 Lepere G, Nowacki M, Serrano V, Gout JF, Guglielmi G, Duharcourt S, Meyer E. 2009. Silencing-
852 associated and meiosis-specific small RNA pathways in *Paramecium tetraurelia*. *Nucleic Acids*
853 *Res* **37**: 903–15.
- 854 Lhuillier-Akakpo M, Frapporti A, Denby Wilkes C, Matelot M, Vervoort M, Sperling L, Duharcourt
855 S. 2014. Local effect of enhancer of zeste-like reveals cooperation of epigenetic and cis-acting
856 determinants for zygotic genome rearrangements. *PLoS Genet* **10**: e1004665.
- 857 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000
858 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and
859 SAMtools. *Bioinformatics* **25**: 2078–2079.
- 860 Lilly MA, Duronio RJ. 2005. New insights into cell cycle control from the *Drosophila* endocycle.
861 *Oncogene* **24**: 2765–2775.
- 862 Maliszewska-Olejniczak K, Gruchota J, Gromadka R, Denby Wilkes C, Arnaiz O, Mathy N,
863 Duharcourt S, Betermier M, Nowak JK. 2015. TFIIS-Dependent Non-coding Transcription
864 Regulates Developmental Genome Rearrangements. *PLoS Genet* **11**: e1005383.
- 865 Marmignon A, Bischerour J, Silve A, Fojcik C, Dubois E, Arnaiz O, Kapusta A, Malinsky S,

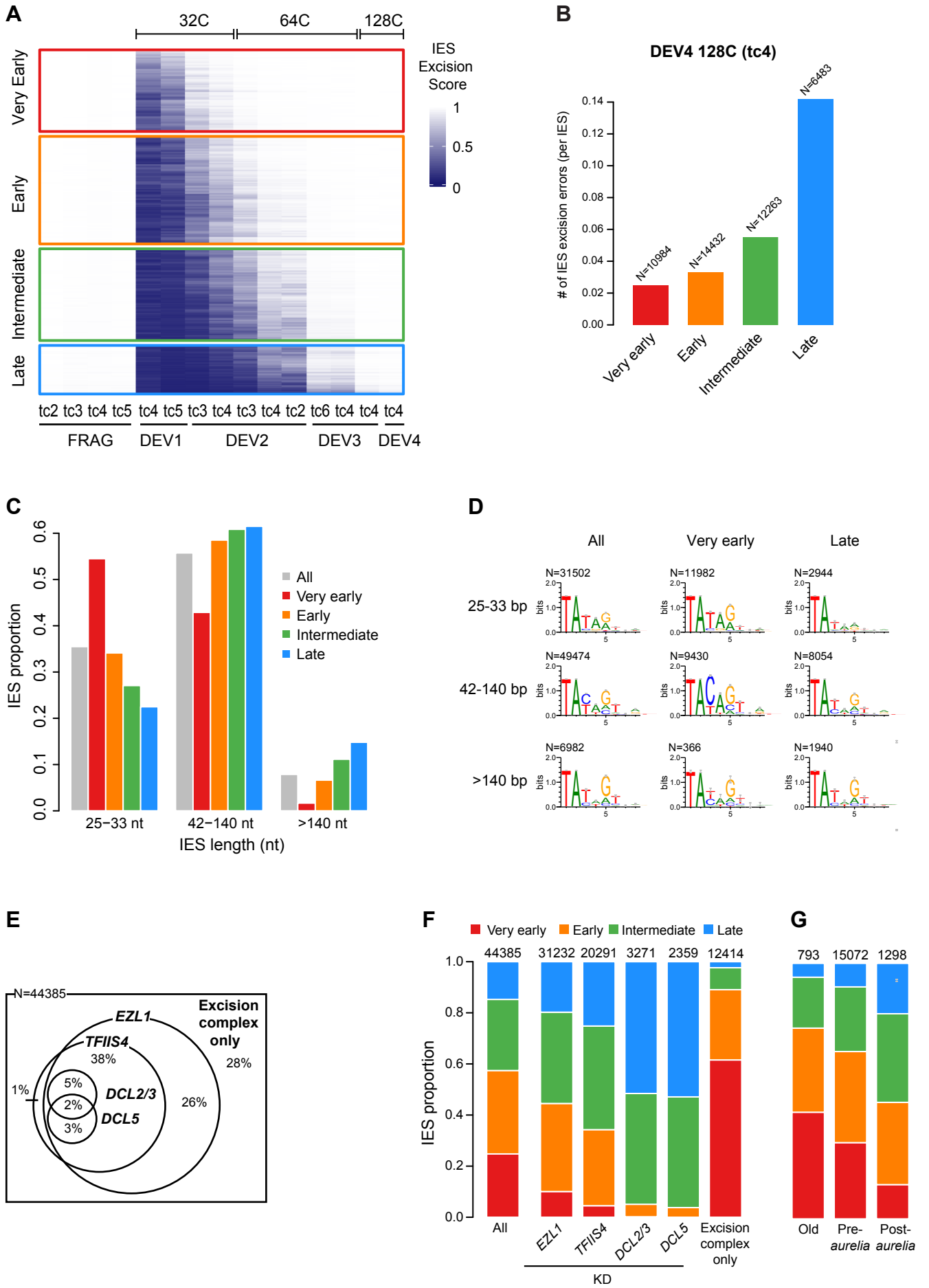
- 866 Betermier M. 2014. Ku-mediated coupling of DNA cleavage and repair during programmed
867 genome rearrangements in the ciliate *Paramecium tetraurelia*. *PLoS Genet* **10**: e1004552.
- 868 Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads.
869 *EMBnet.journal* **17**: 10–12.
- 870 Mayer KM, Forney JD. 1999. A Mutation in the Flanking 5'-TA-3' Dinucleotide Prevents Excision
871 of an Internal Eliminated Sequence From the *Paramecium tetraurelia* Genome. *Genetics* **151**:
872 597–604.
- 873 Mayer KM, Mikami K, Forney JD. 1998. A Mutation in *Paramecium tetraurelia* Reveals Functional
874 and Structural Features of Developmentally Excised DNA Elements. *Genetics* **148**: 139–149.
- 875 Miró-Pina C, Charmant O, Kawaguchi T, Holoch D, Michaud A, Cohen I, Humbert A, Jaszczyszyn
876 Y, Chevreux G, Del Maestro L, et al. 2022. *Paramecium* Polycomb repressive complex 2
877 physically interacts with the small RNA-binding PIWI protein to repress transposable elements.
878 *Developmental Cell*. <https://www.sciencedirect.com/science/article/pii/S1534580722002088>
879 (Accessed April 22, 2022).
- 880 Preer JR. 1976. Quantitative predictions of random segregation models of the ciliate macronucleus.
881 *Genet Res* **27**: 227–238.
- 882 Prescott DM. 1994. The DNA of ciliated protozoa. *Microbiol Rev* **58**: 233–267.
- 883 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
884 *Bioinformatics* **26**: 841–842.
- 885 R Core Team. 2021. R: A Language and Environment for Statistical Computing. [https://www.r-](https://www.r-project.org/)
886 [project.org/](https://www.r-project.org/) (Accessed November 9, 2022).
- 887 Sandoval PY, Swart EC, Arambasic M, Nowacki M. 2014. Functional diversification of Dicer-like
888 proteins and small RNAs required for genome sculpting. *Dev Cell* **28**: 174–88.
- 889 Sardo L, Lin A, Khakhina S, Beckman L, Ricon L, Elbezanti W, Jaison T, Vishwasrao H, Shroff H,
890 Janetopoulos C, et al. 2017. Real-time visualization of chromatin modification in isolated
891 nuclei. *J Cell Sci* **130**: 2926–2940.
- 892 Schleif R. 1992. DNA looping. *Annu Rev Biochem* **61**: 199–223.
- 893 Sellis D, Guerin F, Arnaiz O, Pett W, Lerat E, Boggetto N, Krenek S, Berendonk T, Couloux A,
894 Aury JM, et al. 2021. Massive colonization of protein-coding exons by selfish genetic elements
895 in *Paramecium* germline genomes. *PLoS Biol* **19**: e3001309.
- 896 Singh A, Maurer-Alcalá XX, Solberg T, Gisler S, Ignarski M, Swart EC, Nowacki M. 2022. RNA-
897 mediated nucleosome depletion is required for elimination of transposon-derived DNA.
898 2022.01.04.474918. <https://www.biorxiv.org/content/10.1101/2022.01.04.474918v1> (Accessed
899 May 3, 2022).
- 900 Singh DP, Saudemont B, Guglielmi G, Arnaiz O, Goût J-F, Prajer M, Potekhin A, Przybòs E,
901 Aubusson-Fleury A, Bhullar S, et al. 2014. Genome-defence small RNAs exapted for epigenetic
902 mating-type inheritance. *Nature* **509**: 447–452.
- 903 Smith JJ, Baker C, Eichler EE, Amemiya CT. 2012. Genetic Consequences of Programmed Genome
904 Rearrangement. *Current Biology* **22**: 1524–1529.
- 905 Swart EC, Denby Wilkes C, Sandoval PY, Hoehener C, Singh A, furrer DI, Arambasic M, ignarski
906 M, Nowacki M. 2017. Identification and analysis of functional associations among natural
907 eukaryotic genome editing components [version 1; peer review: 1 approved, 1 approved with
908 reservations]. *F1000Research* **6**.

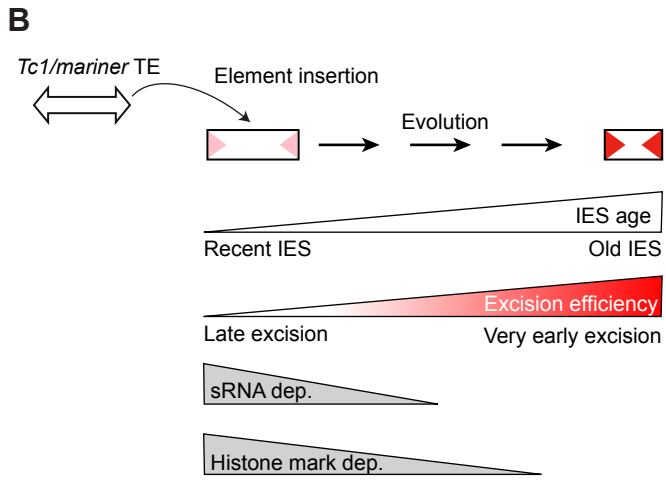
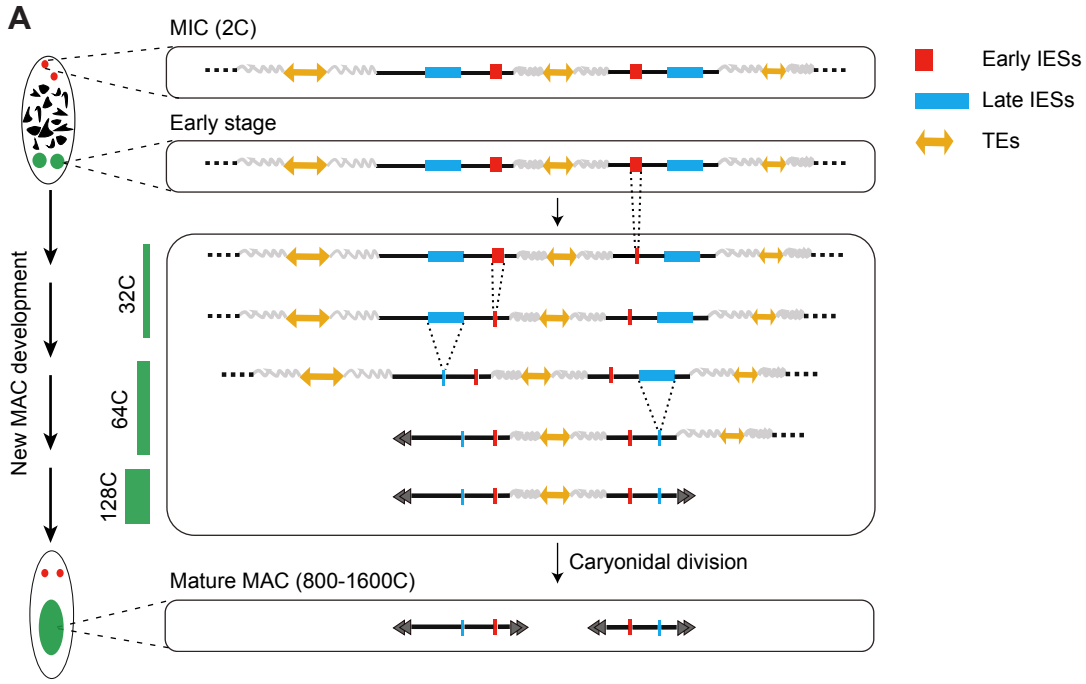
- 909 Swart EC, Wilkes CD, Sandoval PY, Arambasic M, Sperling L, Nowacki M. 2014. Genome-wide
910 analysis of genetic and epigenetic control of programmed DNA deletion. *Nucleic Acids Res* **42**:
911 8970–8983.
- 912 Tabara H, Sarkissian M, Kelly WG, Fleenor J, Grishok A, Timmons L, Fire A, Mello CC. 1999. The
913 rde-1 gene, RNA interference, and transposon silencing in *C. elegans*. *Cell* **99**: 123–32.
- 914 Torgasheva AA, Malinovskaya LP, Zadesenets KS, Karamysheva TV, Kizilova EA, Akberdina EA,
915 Pristyazhnyuk IE, Shnaider EP, Volodkina VA, Saifitdinova AF, et al. 2019. Germline-
916 restricted chromosome (GRC) is widespread among songbirds. *Proceedings of the National*
917 *Academy of Sciences* **116**: 11845–11850.
- 918 Wang C, Solberg T, Maurer-Alcalá XX, Swart EC, Gao F, Nowacki M. 2022. A small RNA-guided
919 PRC2 complex eliminates DNA as an extreme form of transposon silencing. *Cell Rep* **40**:
920 111263.
- 921 Wang J, Mitreva M, Berriman M, Thorne A, Magrini V, Koutsovoulos G, Kumar S, Blaxter ML,
922 Davis RE. 2012. Silencing of Germline-Expressed Genes by DNA Elimination in Somatic
923 Cells. *Developmental Cell* **23**: 1072–1080.
- 924 Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York
925 <https://ggplot2.tidyverse.org>.
- 926 Zilberman D, Cao X, Jacobsen SE. 2003. ARGONAUTE4 control of locus-specific siRNA
927 accumulation and DNA and histone methylation. *Science* **299**: 716–9.
- 928



A**B****C****D**







Legends to Figures

Figure 1. PgmL1 immunostaining during autogamy. (A) Whole cell immunostaining at different stages of autogamy time-course 1 (tc1). New MACs and fragments are counterstained with propidium iodide (PI). Developing MACs are surrounded by a white dotted line. Scale bar is 5 μ m. Developmental stages (DEV1 to 5) are defined in Methods. (B) Flow cytometry analysis of immunostained nuclei at the DEV1 and DEV2 stages of autogamy time-course 2 (tc2). Following gating of total nuclei (see Supplemental Fig. S1D), the population of new MACs was separated based on their PgmL1 signal. The PI axis is indicative of DNA content. A. U.: arbitrary units in log scale.

Figure 2. IES excision kinetics and endoreplication. (A) Flow cytometry sorting of nuclei during the different stages of an autogamy time-course (tc4). Upper panels: plots of PgmL1 fluorescence intensity (y-axis) versus PI fluorescence intensity (x-axis) for nuclei collected at different developmental stages. Lower panels: histograms of PI-stained nuclei gated in the upper panel. Sorted new MAC peaks are indicated by light green shading. The estimated C-level for each sorted peak is indicated above. For DEV4 nuclei, the whole PgmL1-labeled population was sorted (light green), but the major peak was used for calculation of the C-level. As a control, old MAC fragments were sorted from the DEV1 stage. (B) Distribution of IES Excision Scores (ES) in the different sorted new MAC populations. Samples are named according to the developmental stage (DEV1 to DEV4 from tc4) and the C-level of the sorted population. A schematic representation of the IES⁺ and IES⁻ Illumina sequencing reads that were counted to calculate the ES is presented on the left. An ES of 0 or 1 corresponds to no or complete IES excision, respectively. The black dot is the median and the vertical black line delimitates the second and third quartiles. (C) Flow cytometry sorting of nuclei following

aphidicolin treatment. PI histograms of PgmL1-labelled nuclei are presented for each stage or condition (DEV1, DEV3 DMSO and DEV3 Aphi). The C-level for the indicated peaks was estimated as described in Supplemental Table S1. For each stage, all PgmL1-labelled nuclei were sorted. Old MAC fragments were sorted as a control from the DEV3 DMSO nuclear preparation. The dotted line is indicative of a PI value of 10^3 . (D) ES distribution in the sorted new MAC populations in the aphidicolin time-course. Sample names correspond to the sorted samples shown in C.

Figure 3. Kinetics of precise IES excision and imprecise DNA elimination. (A) Distribution of ESs in all samples. Samples are named and ordered according to developmental stage (DEV1 to DEV4), time-course (tc2, tc3, tc4, tc5, tc6) and C-level (indicated above the plot). Hierarchical clustering of ESs confirmed that samples from the same developmental stages (DEV1 to DEV4) group together (Supplemental Fig. S4C). Inside each developmental stage, for a given C-level, we have ordered the samples using their median ES score. For each time-course, old MAC fragments (FRAG) were also sorted as controls. The black dot is the median and the vertical black line spans the second and third quartiles. (B) TE and IES coverage during autogamy. The mean depth coverage distribution is represented as a boxplot. For each dataset described in A, the grey boxplot shows TE coverage, and the white, IES coverage. The percentage of the MIC genome covered by the sequencing reads is indicated above each pair of boxplots. (C) Abundance of telomere addition sites during autogamy. The schema above the bars illustrates the method for detection of telomere addition sites using the sequencing data. For each dataset, the bar shows the normalized number (per million mapped reads, RPM) of detected telomere addition sites localized at less than 10 nt (black), between 10 and 100 nt (dark grey) and more than 100 nt (light grey) from an IES.

(D) Quantification of IES-IES junctions. All putative molecules resulting from ligation of excised IES ends (See Supplemental Fig. S5A,B) are counted and normalized using sequencing depth. The percentage of IESs involved in at least one IES-IES junction is indicated above the barplot.

Figure 4. Excision timing defines IES classes with different characteristics. (A) Heatmap of ESs for all IESs. IESs are sorted by hierarchical clustering, each row corresponding to one IES. The ES is encoded from 0 (dark blue, no excision) to 1 (white, complete excision). IESs are separated in 4 classes according to their excision profile by *k*-means clustering of their ES (very early: N=10,994, early: N=14,490, intermediate: N=12,353, late: N=6,548). (B) Abundance of IES excision errors in the four excision profile groups counted in the DEV4 128C (tc4) sample. In this analysis, we focused on error types that would be the least impacted by IES length (external, overlap and partial external, see Supplemental Fig. S7A). The number of IESs in each excision profile group is indicated above the bars. (C) IES fraction for IES length categories in the four excision profile groups compared to all IESs. (D) Sequence logos of the 8 bases at IES ends for all IESs and IESs belonging to the very early and late clusters. IESs are grouped in three length categories as described in C. (E) Venn diagram showing how the 44,385 reference IESs are distributed according to their sensitivity to *EZL1*, *TFIIS4*, *DCL2/3* and *DCL5* RNAi with regard to excision. The group "excision complex only" represents IESs that do not depend on any of these factors but do depend upon Pgm. The Venn diagram has been simplified to display only overlaps representing more than 1% of the total number of IESs. (F) IES proportions in the 4 groups of excision profiles for the datasets defined in E. The numbers above the barplots indicate the number of IESs in each dataset. "All" is the random expectation for all IESs. (G) IES proportions in the 4 groups of excision profiles relative to the age of IES insertion during evolution of the *Paramecium*

lineage. Old: insertion predating the divergence between *P. caudatum* and the *P. aurelia* clade. Pre-*aurelia*: insertion before the radiation of the *P. aurelia* complex. Post-*aurelia*: insertion after the radiation of the *P. aurelia* complex (Sellis et al. 2021).

Figure 5. Schematic view of DNA elimination timing in *Paramecium* and model for IES evolution. (A) Relative timing of DNA amplification and PDE during new MAC development. The wavy grey lines stand for imprecisely eliminated sequences. The endoreplication level (C-level) is indicated as a green bar on the left. The black double arrowheads schematize the telomeric ends of MAC chromosomes. At each step of PDE, only one representative copy of the new MAC genome is drawn. (B) Model for evolutionary optimization of IESs. Old IESs have become independent of sRNAs and histone mark deposition for their excision. They have acquired strong sequence information at their ends (red arrowheads), promoting their efficient excision.