Massive colonization of protein-coding exons by selfish genetic elements in *Paramecium* germline genomes

Short title: Massive invasion of *Paramecium* genes by selfish genetic elements

Diamantis Sellis[1#a¶], Frédéric Guérin[2#b¶], Olivier Arnaiz[3], Walker Pett[1], Emmanuelle Lerat[1], Nicole Boggetto[2], Sascha Krenek[4], Thomas Berendonk[4], Arnaud Couloux[5], Jean-Marc Aury[5], Karine Labadie[6], Sophie Malinsky[7,8], Simran Bhullar[7], Eric Meyer[7], Linda Sperling[3], Laurent Duret[1&*], Sandra Duharcourt[2&*]

[1] Université de Lyon, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, F-69100, Villeurbanne, France

[2] Université de Paris, CNRS, Institut Jacques Monod, F-75006 Paris, France

[3] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France.

[4] TU Dresden, Institute of Hydrobiology, Dresden, Germany

[5] Génomique Métabolique, Genoscope, Institut de biologie François Jacob, CEA, CNRS, Université d'Évry, Université Paris-Saclay, F-91000 Evry, France.

[6] Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, F-91000 Evry, France.

[7] Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM, Université PSL, Paris, France.

[8] Université de Paris, F-75006 Paris, France

[#a]Current Address: CosmoTech, 5 Passage du Vercors 69007 Lyon, France
[#b]Current Address: Scipio bioscience, 92120 Montrouge, France

* Corresponding authors
Emails: sandra.duharcourt@ijm.fr; Laurent.Duret@univ-lyon1.fr

¶These authors contributed equally to this work.
&These authors also contributed equally to this work.

Abstract

Ciliates are unicellular eukaryotes with both a germline genome and a somatic genome in the same cytoplasm. The somatic macronucleus (MAC), responsible for gene expression, is not sexually transmitted but develops from a copy of the germline micronucleus (MIC) at each sexual generation. In the MIC genome of *Paramecium tetraurelia*, genes are interrupted by tens of thousands of unique intervening sequences, called Internal Eliminated Sequences (IESs), that have to be precisely excised during the development of the new MAC to restore functional genes. To understand the evolutionary origin of this peculiar genomic architecture, we sequenced the MIC genomes of nine *Paramecium* species (from ~100 Mb in *P. aurelia* species to > 1.5 Gb in *P. caudatum*). We detected several waves of IES gains, both in ancestral and in more recent lineages. While the vast majority of IESs are single-copy in present-day genomes, we identified several families of mobile IESs, including non-autonomous elements acquired via horizontal transfer, that generated tens to thousands of new copies. These observations provide the first direct evidence that transposable elements can account for the massive proliferation of IESs in *Paramecium*. The comparison of IESs of different evolutionary ages indicates that, over time, IESs shorten and diverge rapidly in sequence while they acquire features that allow them to be more efficiently excised. We nevertheless identified rare cases of IESs that are under strong purifying selection across the *aurelia* clade. The cases examined contain or overlap cellular genes that are inactivated by excision during development, suggesting conserved regulatory mechanisms. Similar to the evolution of introns in eukaryotes, the evolution of *Paramecium* IESs highlights the major role played by selfish genetic elements in shaping the complexity of genome architecture and gene expression.

57 ## Introduction

58

59 In multicellular organisms, the division of labor between transmission and expression of the
60 genome is achieved by separation of germline and somatic cells. Such a division is also
61 observed in some unicellular eukaryotes, including ciliates [1]. The ciliate *Paramecium*
62 *tetraurelia* separates germline and somatic functions into distinct nuclei in the same cell.
63 Somatic functions are supported by the highly polyploid macronucleus (MAC) that is
64 streamlined for gene expression and destroyed at each sexual cycle. Germline functions are
65 ensured by two small, diploid micronuclei (MIC) that are transcriptionally silent during
66 vegetative growth. During sexual events, the MICs undergo meiosis and transmit the germline
67 genome to the zygotic nucleus. New MICs and new MACs differentiate from mitotic copies of
68 the zygotic nucleus. MAC differentiation involves massive and reproducible DNA elimination
69 events (for review: [2,3]). In addition to the variable elimination of large regions containing
70 repeats, ~45,000 unique, short, interspersed Internal Eliminated Sequences (IESs) are precisely
71 removed from intergenic and coding regions [4,5]. Precise excision of IESs at the nucleotide
72 level is essential to restore functional cellular genes, since 80% of the IESs are inserted within
73 protein-coding genes, and about half of the ~40,000 genes are interrupted by IESs. IESs are
74 invariably bounded by two 5'-TA-3' dinucleotides, one of which is left at the junction in the
75 MAC genome after excision. IES excision in the developing MAC is initiated by DNA double-
76 strand breaks at IES ends by the endonuclease PiggyMac (Pgm) assisted by other proteins,
77 which are likely part of the excision machinery or interact with it [6–9].

78

79 Despite significant progress in characterization of the mechanisms underlying IES elimination,
80 the evolutionary origin of IESs remains mysterious. On the basis of sequence similarities
81 between the consensus found adjacent to the TA dinucleotide at IES ends and the extremities

82 of DNA transposons from the IS630-Tc1-mariner (ITm) superfamily, Klobutcher and Herrick
83 hypothesized that IESs might be degenerated remnants of transposable elements (TEs)[10,11].
84 This hypothesis was further substantiated by the discovery that the endonuclease responsible
85 for IES excision in *P. tetraurelia* is encoded by a domesticated PiggyBac transposase [6],
86 assisted by a related family of catalytically inactive transposases [7]. All-by-all sequence
87 comparison of the *P. tetraurelia* 45,000 IESs and of their flanking sequences identified eight
88 families of "mobile IESs" (2 to 6 copies), *i.e.* homologous IESs inserted at non-homologous
89 sites in the genome [4]. One such family (with 6 copies) was found similar to the Terminal
90 Inverted Repeats of *Thon*, a DNA transposon of the ITm superfamily, indicating that some IESs
91 derive from TEs [4]. These cases provided support to the notion that at least some IESs have
92 derived from recently mobilized elements. However, the rather small number of mobile IESs
93 detected (23 copies out of 45,000 IESs) suggested a limited activity of transposable IESs in the
94 recent evolutionary history of the *P. tetraurelia* lineage [4]. There is also evidence that some
95 IESs originated from MAC sequences, as described for instance for the IESs involved in mating
96 type determination in several species [12,13]. The extent to which the 45,000 IESs detected in
97 *P. tetraurelia* derive from TEs or from MAC sequences therefore remained unclear.
98
99 In order to gain insight into the evolutionary origin of IESs in the *Paramecium* lineage, we
100 adopted a comparative genomic approach. *P. tetraurelia* belongs to the *Paramecium aurelia*
101 group of species that comprises a score of morphologically similar yet genetically isolated
102 species [14–17] (see [13] for a detailed description of the species). Here, we sequenced the
103 germline MIC genomes of eight *P. aurelia* species and one outgroup (*P. caudatum*), compared
104 the IES repertoire across these nine species, and analyzed the evolutionary trajectories of IESs
105 in the *Paramecium* lineage.
106

107 **Results**

108

109 Sequencing of somatic and germline genomes in nine *Paramecium* species: gigantic germline

110 genome in *P. caudatum*

111 In order to determine the origin and evolution of IESs in the *Paramecium* lineage, we sequenced

112 the germline MIC genome and the somatic MAC genome of several *Paramecium* species. We

113 selected 8 species from the *Paramecium aurelia* complex and one outgroup species, *P.*

114 *caudatum*, which diverged from the *aurelia* complex before the two most recent *Paramecium*

115 whole genome duplications [15]. To sequence the germline MIC genome, we purified the

116 germline nuclei (MICs) of each species using a flow cytometry procedure that we previously

117 developed for *P. tetraurelia* [5] (S1 Fig), and used paired-end Illumina sequencing (see

118 Materials and Methods and S1 Table).

119

120

Species (strain)	MIC genome		MAC-destined regions			
	Size (Mb) (a)	Sequencing Depth (b)	Size (Mb) (c)	Nb. of protein genes	Nb. of IESs (d)	IES density (per kb) (e)
P. tetraurelia (51)	108	31	72	40,460	44,128	0.62
P. octaurelia (138)	108	56	72	44,398	44,509	0.61
P. biaurelia (V1-4)	119	58	74	40,261	45,384	0.65
P. tredecaurelia (209)	127	71	65	36,179	42,275	0.66
P. pentaurelia (87)	112	194	72	41,676	42,686	0.57
P. primaurelia (AZ9-3)	114	65	73	42,619	43,766	0.59
P. sonneborni (ATCC 30995)	286	13	82	49,951	60,198 (~85,000)	1.05
P. sexaurelia (AZ8-4)	123	141	68	36,094	47,002	0.70

P. caudatum (My43c3d)	1,300	13	30	18,673	8,762 (~15,000)	0.47

121

122 **Table 1. Characteristics of analyzed genomes.**

123 Species are ordered according to the phylogeny (Fig 1). (a) MIC genome size (in Mb) was estimated
124 based on k-mer counts (see S1 Table for additional estimates based on flow cytometry). (b) Sequencing
125 depth in MIC-specific regions estimated from the distribution of k-mer counts (S2 Fig). (c) Size of
126 constitutive MAC genome assembly. (d) The sensitivity of IES detection was limited in *P. caudatum*
127 and *P. sonneborni* because of the relatively low sequencing depth of their MIC genome. Estimates of
128 total number of IESs, based the IES density observed in regions with sufficient read depth (see e), are
129 indicated in parenthesis for these two species. (e) IES density measured in MAC-destined sequences,
130 after exclusion of regions with insufficient MIC read depth (< 15X).

131

132 To estimate the size of the MIC genomes, we employed two distinct approaches. First, we used
133 the MIC-enriched preparations from *Paramecium* cultures to yield values for DNA quantity in
134 the MICs by flow cytometry (see Materials and Methods and S1 Data). The estimated MIC
135 genome sizes are within a similar range (140-173 Mb) for the *P. aurelia* species, except for *P.*
136 *sonneborni* (S1 Table), which was estimated to be roughly the double of the others. The second,
137 independent approach for genome size estimations was based on the sequence reads themselves
138 and used the k-mer method described in [18,19]. The estimated MIC genome sizes were
139 comprised between 108 Mb to 123 Mb for the *aurelia* species, with a considerably larger MIC
140 genome (283 Mb) for *P. sonneborni* (Table 1). While the values obtained using the flow
141 cytometry method were greater than those with the k-mer method, the estimated MIC genome
142 sizes were within a similar range for both methods (S3 Fig).

143

144 With both methods, the estimated MIC genome size of *P. caudatum* strain My43c3d was the
145 largest among the species analyzed (app. 1,300-1,600 Mb). To confirm this observation, we
146 estimated the genome size of 9 additional strains belonging to the two major clades A and B
147 described in the *caudatum* lineage, as well as another divergent strain [20]. The data confirmed
148 that the MIC genome size in the *caudatum* lineage is far bigger than that in the *aurelia* lineage
149 and revealed great variations of genome size among the different strains (from 1,600 Mb to
150 5,500 Mb), even within the same clade (S1 Data). To investigate the composition of the gigantic
151 *caudatum* genomes, we searched for the presence of repeats in the MIC sequence reads of strain
152 My43c3d. We identified two major satellite repeats, Sat1 and Sat2 (332 bp and 449 bp,
153 respectively), which represent 42% and 29% respectively of the MIC genome (S4 Fig). Both
154 Sat1 and Sat2 repeats were detected in the *P. caudatum* strains of the clade B, to which the
155 strain My43c3d belongs, but not in the other *P. caudatum* strains (S4 Fig). Thus, the repeats are
156 not shared by all *P. caudatum* strains and most likely invaded the MIC genome after the
157 divergence between clades A and B.
158
159 The MAC genome was sequenced for four *aurelia* species that had not been sequenced
160 previously (S2 Table). We defined the '*constitutive'* MAC genome as the DNA sequences
161 retained in all MAC copies (see Materials and Methods and S2 Table). The size of the
162 constitutive MAC genome assembly was similar among *P. aurelia* species (66-73 Mb) with a
163 noticeably larger size for *P. sonneborni* (83 Mb) (Table 1). The number of protein coding genes
164 follows a similar distribution (36,179 to 42,619) in *aurelia* species, with a larger number of
165 genes in *P. sonneborni* (Table 1). This contrasts with the much smaller MAC genome size and
166 number of genes of the outgroup *P. caudatum* [15].
167

168 In conclusion, the eight species of the *aurelia* complex that we analyzed share similar genome
169 characteristics, with a MIC genome of ~110-160 Mb, 50-70% of which is retained during MAC
170 development (~70-80 Mb). The only notable exception is *P. sonneborni*, with a 300-400 Mb
171 MIC genome, of which about 25% is retained in its MAC. The MIC genome of the outgroup
172 *P. caudatum* is much larger (~1,300-1,600 Mb). Only 2% of MIC sequences are retained in the
173 MAC of *P. caudatum* strain My43c3d and 83% of the MIC-specific sequences consist of
174 repeated DNA (S4 Fig).
175

176 ## IES repertoire

177 IESs were identified by comparing MIC sequence reads to the MAC genome assembly (see
178 Materials and Methods; [4,21]). Overall, the number of detected IESs in MAC-destined
179 sequences is similar across *Paramecium* species (~42,000-47,000 IESs), with the exception of
180 *P. sonneborni* (~60,000 IESs) and *P. caudatum* (~9,000 IESs). It should be noted that the
181 sensitivity of IES detection depends on sequencing depth (S5 Fig). For most species, we expect
182 that >99% of IESs have been detected, except in *P. caudatum*, and *P. sonneborni*, where the
183 sequencing depth was limited, due to the unexpected large size of their MIC genome (Table 1).
184 To circumvent this issue, we compared the IES density across species by taking into account
185 only IESs annotated in regions with at least 15X depth of MIC sequence reads mapped onto the
186 MAC assembled genome (Table 1): in *P. caudatum*, the density of detected IES sites in MAC-
187 destined regions (0.5 IESs per kb) is only slightly lower than in other species (~0.6 IESs per
188 kb). This suggests that the genome of *P. caudatum* probably contains about 15,000 IESs in its
189 MAC-destined regions.
190 Our approach is designed to identify IESs only if they are present within loci retained in the
191 MAC. Hence, IESs located in MIC-specific regions (e.g., IESs nested within other IESs
192 [22,23]) remain undetected. Interestingly, in four species whose MAC genome was sequenced

193 at very high depth *(P. octaurelia, P. primaurelia, P. pentaurelia* and *P. sonneborni*), the initial
194 MAC genome assemblies included 10 to 16 Mb of MAC-variable regions, which correspond
195 to DNA sequences that are not completely eliminated and instead, are retained in a small
196 fraction of MAC copies (see Materials and Methods, S6 Fig and S2 Table). We identified many
197 IESs in these regions, at a density (0.4 to 0.5 IESs per kb, S2 Table) nearly as high as in MAC-
198 destined regions (Table 1). This suggests that in addition to IESs located in MAC-destined
199 regions, many other IESs are present within MIC-specific regions.
200
201 In all species, the vast majority of IESs in MAC-destined regions (73% to 81%) are located in
202 protein-coding exons and ~5% are located in introns. Overall, there is a slight enrichment of
203 IESs within genes (S3 Table). This enrichment is not true for all gene categories. In particular,
204 we observed a depletion of IESs in highly expressed genes: on average, the IES density in the
205 top 10% most expressed genes is 37% lower than in the bottom 10% (S7 Fig). This pattern,
206 consistent with previous observations in *P. tetraurelia*, suggests that IES insertions are counter-
207 selected in highly expressed genes [4].
208

209 ## Age distribution of IESs

210 In order to explore the origin and evolution of IESs, we resolved the phylogenetic relationship
211 among the sequenced species. To do so, we classified all protein sequences into families
212 (N=13,617 gene families) and inferred the species phylogeny using the subset of 1,061 gene
213 families containing one single sequence from each species. In agreement with previous reports
214 [16,24], we found strong support for a division of the *aurelia* complex in two subclades
215 (hereafter referred to as subclades A and B), separating *P. sonneborni* and *P. sexaurelia* from
216 the other *aurelia* species (Fig 1). We then used this species phylogeny to identify gene

217 duplications and speciation events in each of the 13,617 gene families, using the PHYLDOG
218 tree reconciliation method [25].

219 **Fig 1. Dynamics of IES insertion/loss in *Paramecium*.**
220 The species phylogeny was reconstructed from a concatenated alignment of 1,061 single-copy
221 genes. All internal branches are supported by 100% bootstrap values (except branch *:
222 bootstrap support = 83%). The age of IESs located within coding regions was inferred from the
223 pattern of presence/absence within gene family alignments (N=13,617 gene families). Only
224 IESs present within well-aligned regions were included in this analysis. The number of dated
225 IESs and the fraction predicted to be old (predating the divergence between *P. caudatum* and
226 the *P. aurelia* lineages), intermediate (before the radiation of the *P. aurelia* complex) or recent
227 are reported for each species. Rates of IES gain (in red) and loss (in blue) were estimated along
228 each branch using a Bayesian approach. Gain rates are expressed per kb per unit of time (using
229 the branch length – in substitutions per site – as a proxy for time). Loss rates are expressed per
230 IES per unit of time (see S8 Fig for more details on the measure of gain/loss rates). NB:
231 estimates of loss rate along terminal branches of the phylogeny also include false negatives (i.e.
232 IESs that are present but that have not been detected), and hence may be overestimated.

233

234 In order to date events of IES gain or loss, we mapped the position of IES excision sites in
235 multiple alignments of each gene family (nucleic sequence alignments based on protein
236 alignments): IESs located at the exact same position within a codon were assumed to derive
237 from a single ancestral insertion event, whatever their present level of sequence similarity (S9
238 Fig). IESs located at homologous sites are hereafter referred to as 'co-orthologous' IESs. We
239 use the term 'co-orthologous' rather than 'orthologous', because these sets frequently include
240 paralogs, notably as a result of whole genome duplications. To avoid ambiguities due to low
241 quality alignments, we only analyzed IESs present within well-conserved protein-coding

242 regions (which represent from 45% to 51% of IESs located in coding regions; Fig 1). We then
243 used the reconciled gene tree to map events on the species phylogeny and estimate rates of IES
244 gain and loss along each branch of the species tree using a Bayesian approach accounting for
245 IES losses and missing data (see Materials and Methods). In the absence of fossil records, it is
246 impossible to date speciation events (in million years). We therefore used sequence divergence
247 (number of amino-acid substitutions per site) along branches of the phylogeny as a proxy for
248 time.
249 Overall, 10.8% of IESs detected in *aurelia* species predate the divergence from *P. caudatum*
250 (referred to as 'Old' IESs in Fig 1), 79% were gained after the divergence of *P. caudatum*, but
251 before the radiation of the *aurelia* complex ('Intermediate' in Fig 1) and 10.2% are more recent.
252 The rate of IES gain varied widely over time: a burst of insertions occurred in the ancestral
253 branch leading to the *aurelia* clade, followed by a progressive slowdown in most lineages,
254 except in *P. sonneborni* where the rate of IES gain strongly increased again in the recent period
255 (18.8% of IESs detected in *P. sonneborni* are specific to that species). The IES gain rate has
256 remained substantial in *P. sexaurelia* and *P. tredecaurelia*, but has dropped to very low levels
257 in *P. tetraurelia/P.octaurelia* and in *P. pentaurelia/P. primaurelia* lineages, about 20 times
258 lower than in *P. sonneborni* or in the ancestral *aurelia* lineage (Fig 1). The rate of IES loss
259 appears to be more uniform along the phylogeny, with only 2 to 3-fold variation (Fig 1).
260

261 ## Recent waves of mobilization of IESs

262 The episodic bursts of IES gains that we observed in the phylogeny are reminiscent of the
263 dynamics of invasion by TEs. To test the hypothesis that IESs might correspond to TEs, we
264 searched for evidence of mobile IESs, i.e. IES sequences sharing significant sequence similarity
265 but inserted at different (non-homologous) loci. In a first step, we compared all IESs against
266 each other with BLASTN to identify clusters of similar IESs. Consistent with previous

observations made in *P. tetraurelia* [4], we found that the vast majority of IESs correspond to unique sequences in all MIC genomes, but a fraction of IESs are present in multiple copies (Fig 2A). In a second step, all clusters with ≥ 10 interspersed copies were manually inspected, to precisely delineate the boundaries of the repeated element and create a multiple alignment of full-length copies. We then used these representative multiple alignments to perform an exhaustive sequence similarity search based on HMM profiles over the entire IES dataset (see Materials and Methods). Among the hits, we distinguished two categories: 1) cases where the detected copy is located within the IES but does not include the extremities of the IES, and 2) cases where the extremities of the copy correspond precisely to the extremities of the IES. The first category probably corresponds to TEs that were inserted within a pre-existing IES (i.e. nested repeats). The second category corresponds to cases where the transposed element is the IES itself (i.e. mobile IESs). Overall, we detected 24 families with at least 10 copies of mobile IESs, totaling 7,443 copies of mobile IESs (Table 2).

Repeat family	Length (bp)	Nb. of mobile IESs	Number of mobile IESs per species								
			pca	pse	pso	ptr	ppe	ppr	pbi	poc	pte
FAM_2183	233	4153	0	4	3252	897	0	0	0	0	0
FAM_3	290	1783	0	0	0	15	344	321	766	146	191
FAM_2938	765	378	0	7	370	1	0	0	0	0	0
FAM_2317	768	331	0	82	140	109	0	0	0	0	0
FAM_2942	211	110	0	0	110	0	0	0	0	0	0
FAM_2334	214	106	0	17	89	0	0	0	0	0	0
FAM_2321	471	84	1	11	58	10	1	1	2	0	0
FAM_78	50	56	0	0	0	0	34	22	0	0	0
FAM_1402 (TIR *Thon)*	693	53	0	6	8	5	0	0	3	16	15
FAM_1257 (TIR *Merou)*	522	49	0	0	5	5	3	2	7	12	15

FAM_670	46	44	0	0	9	1	2	5	2	24	1
FAM_2649	762	40	0	0	16	24	0	0	0	0	0
FAM_1473	98	33	0	0	4	7	1	0	5	10	6
FAM_51	231	32	0	0	0	0	12	9	11	0	0
FAM_692	93	26	0	0	4	13	5	4	0	0	0
FAM_1294 (*Baudroie*)	1706	26	0	0	0	0	1	2	18	4	1
FAM_2314 (DDE)	3421	24	0	20	2	2	0	0	0	0	0
FAM_2802	32	22	0	0	2	20	0	0	0	0	0
FAM_3194	230	20	20	0	0	0	0	0	0	0	0
FAM_837	50	18	0	0	0	0	14	4	0	0	0
FAM_1165	77	15	0	0	0	0	1	1	0	12	1
FAM_2936	223	15	0	0	15	0	0	0	0	0	0
FAM_1259	231	14	0	0	0	0	1	0	13	0	0
FAM_3023	350	11	0	0	11	0	0	0	0	0	0
Total		7443	21	147	4095	1109	419	371	827	224	230

Table 2. Taxonomic distribution of mobile IESs.

Copies of mobile IESs were searched among IES sequences (see Materials and Methods). Detected copies were divided in two categories: nested copies (i.e. copies inserted within an IES, but not including the extremities of the IES) and *bona fide* mobile IESs (i.e. copies whose extremities correspond to the extremities of the IES). This table lists all families for which at least one species contains ≥ 10 copies of *bona fide* mobile IESs in its genome (see S4 Table for information on nested copies). Species codes: pso: *P. sonneborni*, ptr: *P. tredecaurelia*, pte: *P. tetraurelia*, pbi: *P. biaurelia*, poc: *P. octaurelia*, pse: *P. sexaurelia*, ppr: *P. primaurelia*, ppe: *P. pentaurelia*, pca: *P. caudatum.*

Four of these mobile IESs present homology with DNA transposons of the ITm superfamily previously identified in *P. tetraurelia* [4,5] (Table 2). FAM_2314 (3.4 kb) includes an intact open reading frame (ORF) encoding a DDE transposase. FAM_1294 (1.7 kb) is homologous

294 to *Baudroie*, a composite Tc1-mariner element, and includes an ORF with similarity to tyrosine-
295 type recombinases. FAM_1402 (0.7 kb) and FAM_1257 (0.5 kb) correspond to non-
296 autonomous elements, homologous to the terminal inverted repeats (TIR) of *Thon* and *Merou*
297 respectively. The other families of mobile IESs do not match with any known TEs. Their
298 relatively short lengths (32 bp to 765 bp) and the absence of similarity with any known protein,
299 indicate that they most probably correspond to non-autonomous elements, mobilized by
300 transposases expressed from active TEs.
301 The genomic distribution of mobile IESs within MAC-destined regions is similar to that of
302 other IESs: most of the families are predominantly located within protein-coding regions
303 (which represent ~70% of the MAC genome) (S4 Table). The only notable exceptions are
304 FAM_1257, FAM_1402 and FAM_78 elements, which are under-represented within genes (S4
305 Table). In particular, FAM_78 elements are exclusively found in intergenic regions.
306
307 As explained previously, it is possible to date insertions for the subset of IESs located within
308 well-conserved protein-coding regions. The vast majority (97.5%) of mobile IES copies that
309 can be dated correspond to recent insertions (as compared to only 9.5% of recent insertions for
310 the other IESs). FAM_3 is present in all genomes of the subclade A (Table 2), and 94% of dated
311 insertions are shared by at least two species, which indicates that this element has been very
312 active at the beginning of the radiation of this clade. For the other families of mobile IESs, more
313 than 97% of insertion loci are species-specific. Thus, all the families of mobile IESs that we
314 detected are relatively recent, posterior to the radiation of the *aurelia* clade. This most probably
315 reflects the fact that more ancient families are difficult to recognize, because of the rapid
316 divergence of IES sequences.
317

318 In agreement with previous analyses [4], we detected few cases of recent mobilization of IESs
319 in *P. tetraurelia* (although we detected more copies than reported by [4], owing to the higher
320 sensitivity of the sequence similarity search protocol used here). However, several other species
321 contain IES families with a high number of recently inserted copies (Table 2). The largest
322 family (FAM_2183) corresponds to a 233 bp-long non-autonomous element, for which we
323 detected a total of 4,126 copies in *P. sonneborni*, and 1,089 copies in *P. tredecaurelia*, among
324 which 20% are nested within IESs and 80% correspond to mobile IESs (Table 2, Fig 2B-D).
325 The high level of sequence similarity between all copies (average pairwise identity=72%; Fig
326 2) suggests that both species have been invaded recently by this mobile element. To assess
327 whether some insertions are ancestral to these two species, we analyzed copies inserted in well-
328 conserved coding regions (i.e. for which it is possible to check whether they are co-orthologous
329 or not; N=1,973 copies). We identified only two FAM_2183 elements inserted at co-
330 orthologous sites in both species. Thus, the quasi-totality of FAM_2183 copies result from
331 independent waves of insertion in the two lineages. Interestingly, FAM_2183 is absent from
332 other *Paramecium* species (except 4 copies in *P. sexaurelia*). It is important to note that *P.*
333 *sonneborni* and *P. tredecaurelia* belong to two distantly related subclades of the *aurelia*
334 complex (Fig 1). This patchy phylogenetic distribution therefore suggests that FAM_2183 has
335 been subject to horizontal transfers between *P. sonneborni* and *P. tredecaurelia* lineages. The
336 alternative hypothesis of vertical inheritance of the family since the last common ancestor of
337 the *aurelia* complex would imply that all copies of the family have been lost independently in
338 at least three lineages of subclade A. Furthermore, the phylogeny of this family is mainly star-
339 like, with *P. sonneborni* and *P. tredecaurelia* copies interspersed across the entire tree (Fig 2B).
340 This topology is incompatible with a scenario of vertical inheritance, in which copies from each
341 species would have been expected to form two distant clades. Interestingly, the tree topology
342 indicates that several events of horizontal transfer, not just one, occurred between these two

343 lineages. This finding is corroborated by four other families (FAM_2317, FAM_2321,
344 FAM_2649, FAM_2802) that are shared specifically by *P. tredecaurelia* and the *P.*
345 *sonneborni/P. sexaurelia* clade. These multiple events of horizontal transfer suggest an
346 important genetic flux between these two lineages. And indeed, besides these mobile IESs, we
347 also found many large genomic segments (up to 200 kb) that appear to have been introgressed
348 into MIC-specific regions of *P. sonneborni* genome, from a lineage closely related to *P.*
349 *tredecaurelia* (these results will be described in detail in another article). The hypothesis that
350 we propose is that these introgressed segments brought some active transposable elements (both
351 autonomous or non-autonomous), including some mobile IESs, which then were able to
352 proliferate within the *P. sonneborni* genome. The finding of two FAM_2183 copies present at
353 co-orthologous sites is *a priori* unexpected under this scenario. However, they both correspond
354 to nested copies. Thus, one possible explanation is that, by chance, these copies were inserted
355 independently in both lineages within pre-existing co-orthologous IESs. An alternative
356 hypothesis is that both nested insertions might have been transferred between species through
357 the introgression of larger genomic segments. Like most species of the *aurelia* complex, *P.*
358 *sonneborni* and *P. tredecaurelia* have a worldwide geographic distribution [26,27]. Our
359 observations indicate that despite the strong reproductive isolation between extant species [26],
360 horizontal transfers did occur recently between these genetically very distant lineages.
361
362
363 **Fig 2. Proportion of IESs corresponding to repeated sequences, and phylogenetic**
364 **analysis of the largest family of mobile IESs** (A) The vast majority of IESs correspond to
365 unique sequences, but a fraction of IESs are present in multicopies. For each species, all IESs
366 were compared against each other with BLASTN (with an E-value threshold of 10^{-5}). The
367 distribution of the number of BLAST hits per IES (excluding self-hits) is displayed for each

species. (B) The largest family of multicopy IESs (FAM_2183) includes 4,153 copies. We present here a phylogenetic tree of a subset of sequences (200 IESs from *P. tredecaurelia* in black, and 200 from *P. sonneborni* in red), randomly sampled from the entire FAM_2183 alignment (computed with PhyML [28]). The tree topology is mainly star-like, which indicates that most copies derive from several bursts of insertions. (C) Length distribution of the 4,153 FAM_2183 mobile IESs. (D) Sequence logo [29], based on the alignment of the entire FAM_2183 family. The information content reported in the logo was computed relative to the base composition of IESs, to take into account their high AT-content. All copies present a high level of sequence similarity (average pair-wise identity 72%) throughout their entire length, not just at their ends.

IES excision mechanism varies with IES age

Like any biological process, the excision of IESs during new MAC development is not 100% efficient [4,30]. For instance, the IES retention rate in *P. tetraurelia* MAC chromosomes is on average 0.8% in wild-type cells [31]. We observed that a substantial fraction of IESs have a much lower excision efficiency. In all *Paramecium* species, the proportion of 'weak' IESs (defined as IESs with more than 10% retention in wild-type cells) differs strongly among genomic compartments: from 0.7% on average for IESs located within genes (introns or exons), to 5.4% for IESs in intergenic regions (S10 Fig). This difference probably results from the fact that IESs with low excision efficiency are more deleterious, and therefore more strongly counter-selected, in genes than in intergenic regions. Interestingly, we also observed that within coding regions, the proportion of weak IESs is much higher for newly gained IESs (2.1% on average) than older ones (0.3%) (S10 Fig). This suggests that after a wave of insertions,

393 genomes accumulate changes in cis (within IESs) and/or in trans (in the IES excision
394 machinery) that progressively make these new IESs more efficiently excised, presumably in
395 response to the selective pressure against retention of IESs within coding regions.
396 In *P. tetraurelia*, functional analyses have revealed that different classes of IESs rely on
397 different excision pathways [31–33]. A large subset of IESs (70%) requires the histone H3
398 methyltransferase Ezl1 for excision [31]. A much smaller subset (7%), all among the of *EZL1*-
399 dependent IESs, also requires the Dcl2/3 proteins, which are necessary for the biogenesis of 25
400 nt long scnRNAs [31,34]. The remaining 30% of IESs require neither Ezl1 nor Dcl2/3 to
401 complete excision. Using published IES excision efficiency datasets upon silencing of *EZL1* or
402 *DCL2/3* [31], we found that 92% of newly inserted *P. tetraurelia* IESs are sensitive to Ezl1, as
403 compared to 39% for old ones (Fig 3). Similarly, the proportion of Dcl2/3-dependent IESs
404 varies from 17% for new IESs to 3% for old ones. These observations suggest that newly
405 inserted IESs, like TEs themselves [35], initially depend on histone marks deposited by Ezl1
406 (and to some extent on the scnRNA pathway). Over time, histone marks and scnRNAs become
407 dispensable as IESs gradually acquire features that allow them to be efficiently excised.
408

409 **Fig 3. Old and young IESs rely on different excision pathways**.
410 The subset of IESs whose excision relies on histone marks or scnRNAs have been identified in
411 *P. tetraurelia* by measuring their IES retention score (IRS) upon silencing of *EZL1* and
412 *DCL2/DCL3* [31]. Barplots represent the proportion of sensitive IESs (IRS>10%) according to
413 their age, for each silencing experiment. The age of an IES insertion is defined by the
414 phylogenetic position of the last common ancestor (LCA) of species sharing an IES at the same
415 site (New: *P. tetraurelia*-specific IES; Node *n*: the LCA corresponds to node number *n* in the
416 species phylogeny; Old: the LCA predates the *P. aurelia*/*P. caudatum* divergence). The

417 proportion of sensitive IESs among new IESs was compared to that of older ones by a Chi-
418 squared test (*: p-value < 0.05; **: p-value < 1e-3; ***: p-value < 1e-6).

419

420 We also compared the length of IESs according to their age. As previously observed for *P.*
421 *tetraurelia* [4], IESs have a characteristic length distribution with the same ~10 bp periodicity
422 in all *aurelia* species (Fig 4A), likely reflecting structural constraints on the excision process
423 [4,7]. We observed that the length distribution of IESs changes drastically over evolutionary
424 time. For instance, in *P. sonneborni*, *P. tredecaurelia* and *P. tetraurelia*, the proportion of IESs
425 in the first peak of the length distribution (< 35 bp) ranges from 1-10% for new IESs to 81-84%
426 for old ones (Fig 4B), and similar patterns are observed in all other *aurelia* species (S11 Fig).
427 In *P. caudatum*, the overall length distribution is shifted towards shorter IESs (71% in the first
428 peak, compared to 35% in *aurelia*; Fig 4A). This suggests that this lineage has not been subject
429 to IES insertion waves for a long period of time, in agreement with the paucity of recognizable
430 mobile IESs in that genome (Table 2).

431

432 **Fig 4. Length distribution of IESs according to their age**.
433 (A) Comparison of the length distribution of IESs in *P. caudatum* (N=8,172 IESs) and in species
434 from the *aurelia* clade (N=392,082 IESs). The fraction of IESs present within each peak of the
435 distribution is indicated for the first 10 peaks. (B) Comparison of the length distribution of IESs
436 according to their age (for the subset of datable IESs located in coding regions). The age of IES
437 insertions is defined as in Fig 3. Results from other species are presented in S11 Fig.

438

439 ## Genomic distribution of IESs according to their age

440 Because of the rapid divergence of non-coding sequences, it is generally not possible to assess
441 homology among IES insertion sites located in intergenic regions, and hence it is not possible

442 to date them directly. We therefore used the length of IESs as a rough proxy for their age, to
443 investigate their genomic distribution over time. We observed that in all *aurelia* species, long
444 IESs (> 100 bp, presumably young) are uniformly distributed across genomic compartments
445 (introns, coding regions and intergenic regions) (S12 Fig). Conversely, short IESs (<35 bp,
446 presumably older) are enriched in coding regions (on average, 81% of short IESs in coding
447 regions, vs. 70% expected; S12 Fig). This suggests that IESs located within intergenic regions
448 have a shorter lifespan than those located in coding regions.
449

450 ## Evidence that some IESs are functional

451 A large majority of detected IESs predate the divergence of the *aurelia* clade (Fig 1). Because
452 of the rapid evolution of non-coding sequences, co-orthologous IESs from different species are
453 generally too divergent to be recognized by sequence similarity search. Yet the comparison of
454 all sequences against each other revealed several interesting exceptions. Overall, we identified
455 69 clusters of co-orthologous IESs conserved across at least 5 of the 8 species of the *aurelia*
456 clade, and showing at least 70% identity between sequences from subclades A and B (87%
457 identity on average). Each of these 69 IESs correspond to unique sequences (we did not detect
458 any interspersed homologs within genomes, only co-orthologs). These highly conserved IESs
459 are similar to other IESs in terms of length (mean=75 bp) or genomic distribution (79% within
460 protein-coding genes, 21% in intergenic regions). Their high levels of sequence conservation
461 indicate that they are subject to strong selective constraints, and hence that they have a function,
462 beneficial for *Paramecium*. By definition, IESs are absent from the MAC genome so they
463 cannot be expressed in vegetative cells. However, they can potentially be transcribed during
464 the early development of the new MAC, before IES excision occurs [32,36]. To gain insight
465 into their possible functions, we analyzed the transcription of conserved IESs using
466 polyadenylated RNAseq data from autogamy time course experiments in *P. tetraurelia* [37].

467 Among the 56 families of highly conserved IESs present in *P. tetraurelia*, 10 (18%) are
468 transcribed at substantial levels (>1 RPKM) during autogamy (as compared to 0.8% for other
469 IESs) (S5 Table). One of these IESs (~800 bp-long) contains a gene encoding a putative DNA-
470 binding protein, well conserved in all species of the *aurelia* clade and expressed at high levels
471 during the early stages of autogamy (Fig 5A,C). Several paralogs of this gene are present in the
472 MAC genome, but it is the only member of this family to be located within an IES. We
473 hypothesize that this copy might have originated by retrotransposition of one of the MAC
474 paralogs, and that its insertion within a pre-existing IES might have been exapted, as a way to
475 restrict its expression specifically to early MAC development.
476

477 **Fig 5. Examples of highly conserved co-orthologous IESs.** (A) Cluster 4968 corresponds to
478 a set of co-orthologous IESs that contain a gene (blue), oriented in the opposite direction with
479 respect to the gene (blue) in which the IES is inserted (yellow). The IES-embedded gene
480 encodes a protein belonging to a family of CENP-B DNA-binding proteins, and is likely of
481 cellular origin. RNA-Seq data (shown here for the *Paramecium tetraurelia* member of this
482 cluster: IES.PTET.51.1.16.324097) show that this gene is expressed at meiosis, with mRNA
483 steady state levels strongly decreasing before most genome-wide IES excision occurs [37]. **(**B)
484 Cluster 9405 corresponds to a set of co-orthologous IESs that contain the promoter and the start
485 codon of a gene (blue) encoding a protein of unknown function. The IES is drawn in yellow.
486 RNA-Seq data (shown here for one *Paramecium tetraurelia* member of this cluster:
487 IES.PTET.51.1.28.104046) indicate that this gene is expressed at the time genome
488 rearrangements occur, with a peak in steady-state mRNA level at T11, when IES excision peaks
489 [37]. (D) Nucleotide alignment of co-orthologous IESs from cluster 9405. The ATG start codon
490 and first few amino acids encoded by the IES are shown below the alignment. (C) Alignment
491 of the proteins encoded by IESs from cluster 4968. [IES IDs in the alignments, from top to

492 bottom (C): IES.PPENT.87.1.080.374281, IES.PPRIM.AZ9-3.1.072.352290,
493 IES.PBIA.V1_4.1.0053.93462.rc, IES.POCT.138.1.076.348969, IES.PTET.51.1.61.348482,
494 IES.PTRED.209.1.7180000129326.89420.rc, IES.PTRED.209.1.7180000129337.349035,
495 IES.PBIA.V1_4.1.0022.293968, IES.PPENT.87.1.047.444829, IES.PPRIM.AZ9-
496 3.1.042.468625, IES.PSEX.AZ8_4.1.047.365473, IES.PSEX.AZ8_4.1.067.78433.rc,
497 IES.PSON.ATCC_30995.1.070.107189.rc, IES.POCT.138.1.047.443628,
498 IES.PTET.51.1.28.104046.rc. (D): IES.PSON.ATCC_30995.1.014.460171,
499 IES.PSEX.AZ8_4.1.004.493121, IES.PTRED.209.1.7180000129358.256806,
500 IES.PTET.51.1.16.324097, IES.POCT.138.1.015.320413, IES.PBIA.V1_4.1.0089.86544,
501 IES.PPENT.87.1.022.397764, IES.PPRIM.AZ9-3.1.019.378988.]

502

503 All other highly conserved IESs are much shorter (< 300 bp), most probably too short to encode
504 proteins. But we found examples suggesting that some of them contribute to the regulation of
505 the expression of their host gene. For instance, we identified a conserved IES located at the 5'
506 end of a gene of unknown function, encompassing the transcription start site and the beginning
507 of the first exon (including the 5'UTR and the first 9 codons). The excision of the IES during
508 MAC development leads to the loss of the initiation codon and of the promoter region, and
509 thereby to the silencing of this gene in vegetative cells (Fig 5B,D). These examples illustrate
510 that recruitment of the IES excision machinery during evolution to contribute new functions
511 beneficial for *Paramecium*.

512

513 ## Discussion

514

515 ## A majority of *Paramecium* IES insertions result from the transposition of mobile IESs

516 To explore the evolutionary origin of IESs, we analyzed the MIC genomes of eight species of
517 the *P. aurelia* complex, and of an outgroup species, *P. caudatum*. Unexpectedly, we discovered
518 that the MIC genomes of *P. caudatum* strains are at least one order of magnitude larger than
519 those of *P. aurelia* species (~1,600 to 5,500 Mb vs ~110-160 Mb). The sequencing of *P.*
520 *caudatum* My43c3d revealed that its huge MIC genome size is caused by the amplification of
521 two major satellite repeats, which represent 71% of its MIC-limited genome (S4 Fig). The high
522 variability of genome sizes across the *P. caudatum* lineage makes this clade an attractive model
523 system to study the possible phenotypic consequences of genome size variations within a
524 species.
525
526 All the *Paramecium* MIC genomes we sequenced present a high density of IESs in MAC-
527 destined sequences: from 0.5 IES per kb in *P. caudatum* up to 1 IES per kb in *P. sonneborni*
528 (Table 1). The vast majority of these IESs (83% on average) are located within genes, as
529 expected given the very high gene density in MAC genomes (S3 Table). In *aurelia* species,
530 there are on average 0.95 IESs per protein-coding gene. The IES density varies among genes,
531 but overall, ~50% of the ~40,000 genes contain at least one IES. Moreover, the analysis of
532 MIC-specific regions that are occasionally retained in the MAC (MAC-variable regions)
533 revealed similar IES densities (S2 Table), which suggests that, in addition to IESs located in
534 MAC-destined regions, many other IESs are located within MIC-specific regions.
535
536 To explore the origin and evolution of these tens of thousands of IESs, we sought to identify
537 homologous IESs across the 9 *Paramecium* species. We distinguished two categories of
538 homologs: interspersed homologs (i.e. similar IES sequences located at non-homologous loci)
539 and co-orthologous IESs (i.e. IESs deriving from a single ancestral insertion event at a given

540 genomic site). We first compared all IES against each other to identify sets of homologs based
541 on sequence similarity, within genomes or across species. The vast majority of IESs correspond
542 to sequences that are unique within each genome (Fig 2). We identified only a handful of co-
543 orthologous IES sequences that are conserved in sequence across the *aurelia* complex (see
544 below). Apart from these rare cases, most IES sequences evolve rapidly, as expected for non-
545 coding sequences. It should be noted that species from the *aurelia* complex are genetically quite
546 divergent (S6 Table). For instance, the average synonymous divergence (measured in
547 orthologous protein-coding genes) is 0.95 substitutions/site between species of the A and B
548 subclades. At this evolutionary scale, in the absence of selective pressure, IES homologs are
549 expected to be far too diverged to be recognizable by sequence similarity. Thus, homologs
550 (interspersed or co-orthologs) that can be detected by sequence similarity essentially
551 correspond to IESs whose last common ancestor is relatively recent (more recent than the A
552 and B subclades).
553
554 However, it is also possible to identify co-orthologous IESs based on their shared position
555 within multiple alignments of homologous genes. Thus, for the subset of IESs located in coding
556 regions, we were able to infer rates of IES gain and loss across the species phylogeny (Fig 1).
557 Overall, about 90% of IESs detected in *aurelia* species predate the radiation of that clade, but
558 fewer than 10% are shared with *P. caudatum*. Thus, the vast majority of *aurelia* IESs result
559 from a major wave of IES gains that occurred after the divergence of *P. caudatum,* but before
560 the radiation of the *aurelia* complex. Similarly, 80% of IESs detected in *P. caudatum* are
561 specific to that lineage, which implies that multiple independent events of massive IES
562 invasions occurred during evolution.
563

564 The burst of IES gains at the base of the *aurelia* clade was followed by a progressive slowdown
565 in most species, except in the *P. sonneborni* lineage, which has been subject to a second wave
566 of IES insertions (Fig 1). Interestingly, the comparison of IES sequences revealed thousands of
567 interspersed homologous copies, resulting from the recent and massive mobilization of a small
568 number of IESs. Several families of mobile IESs present homology with known ITm
569 transposons, and some of them encode transposases. But most mobile IESs do not appear to
570 have any protein-coding potential, and therefore likely correspond to non-autonomous
571 elements, whose mobility depends on the expression of active transposons. The number of
572 detectable (recently mobilized) IES copies varies widely across species (Table 2). For instance,
573 mobile IESs have been very active in the *P. sonneborni* lineage (4,095 copies), much more than
574 in its sister lineage, *P. sexaurelia* (147 copies). Thus, mobile IESs account for at least 20% of
575 the difference in IES number between these two species (Fig 1). Given the rapid evolution of
576 IESs, interspersed homologs that can be detected probably represent only the tip of the iceberg
577 of mobile IESs. Overall, we found a strong correlation (R^2=0.86, p=8x10^{-4}) between the number
578 of mobile IES copies detected in each species (Table 2) and the rate of IES gain along
579 corresponding branches of the phylogeny (Fig 1), which suggests that most gains result from
580 transposition.
581
582 Interestingly, the five most active families in *P. tredecaurelia* all show the signature of
583 horizontal transfer with the distantly related *P. sonneborni* lineage (Table 2). This is the case
584 of the largest family that we identified (FAM_2183: 3,252 and 897 copies in each species,
585 respectively; Fig 2). This pattern is reminiscent of the typical life cycle of many DNA
586 transposons: when a new element enters a genome, it is initially very active and produces a
587 wave of insertions. Its activity then progressively slows down, largely because defense
588 mechanisms become more efficient in the host genome. In the long-term, DNA transposons

589 escape extinction only if they can occasionally be transmitted to a new host [38]. Thus, the
590 variation of IES insertion rates that we observed in the *Paramecium* phylogeny fits very well
591 with the dynamics of TEs: rare episodes of massive proliferation (promoted by horizontal
592 transfer of an active copy to a new naive host), followed by progressive slowdown of
593 transposition activity.
594
595 TEs are not the unique source of IES gains. Mutations in MAC-destined regions can generate
596 sequence motifs that are recognized by the IES excision machinery, and thereby create new
597 IESs. There is indeed evidence that cryptic IES signals occasionally trigger the excision of
598 MAC-destined sequences [4,30], and that some IESs originated from MAC-destined sequences
599 [12,13]. However, our results suggest that the vast majority of IESs correspond to
600 unrecognizable fossils of mobile elements – as initially proposed by Klobutcher and Herrick
601 [10,11].
602

603 ### Indirect evidence of mutational burden caused by IES invasions

604 In all *Paramecium* species, we observed a deficit of IESs in highly expressed genes (S7 Fig).
605 As previously reported in *P. tetraurelia* [4], this pattern most probably reflects selective
606 pressure against IES insertions within genes. Indeed, the IES excision machinery (like any other
607 biological machinery) is not 100% efficient: a small fraction of IES copies are retained in the
608 MAC or subject to imprecise excision [30]. Typically, the average IES retention rate in MAC
609 chromosomes is 0.8% in *P. tetraurelia* [31]. For IESs located within genes, such excision errors
610 are expected to have deleterious consequences on fitness, in particular for genes that have to be
611 expressed at high levels [4]. And indeed, in agreement with the hypothesized selective pressure
612 against IESs within genes, we observed that the proportion of 'weak' IESs (i.e. IESs with a
613 relatively high retention frequency) is much lower in genes than in intergenic regions (S10 Fig).

614 Despite their selective cost, weakly deleterious IES insertions can eventually become fixed by
615 random genetic drift. Once fixed, the fitness of the organism will depend on its ability to
616 properly excise the IES during MAC development. Over time, selection should favor the
617 accumulation of substitutions and indels that increase the efficiency of IES excision. Indeed,
618 we did observe that the proportion of weak IESs decreases with their age (S10 Fig).
619 Interestingly, older IESs, which are also shorter, are less dependent on the Ezl1 and Dcl2/3
620 proteins (Fig 3). This suggests that after their insertion, IESs progressively acquire features that
621 make them more efficiently excised, by a pathway that requires neither scnRNAs nor histone
622 marks [31].
623

624 ## Exaptation of the IES excision machinery

625 As mentioned previously, orthologous sequences that evolve neutrally are not expected to
626 display any significant similarity between species of subclades A and B (synonymous
627 divergence ~0.95 substitution per site; S6 Table). Yet, we identified 69 clusters of co-
628 orthologous IESs displaying a high level of sequence conservation across the entire *aurelia*
629 complex (on average 87% identity between the closest pair of homologs from subclades A and
630 B). Their level of conservation implies that these IESs are subject to strong purifying selection,
631 and hence, that they fulfill a function that contributes to the fitness of *Paramecium*. Notably,
632 we identified one IES that contains a protein-coding gene (Fig 5A,C). This gene is expressed
633 during the early stages of autogamy, likely from the new developing MAC, before IES excision
634 (Fig 5A). Interestingly, 18% of the conserved IESs are transcribed during autogamy (as
635 compared to 0.8% for other IESs). Most conserved IESs are too short to encode proteins, but
636 they may contribute to gene regulation (e.g. Fig 5B,D). Given the enrichment of conserved IESs
637 in genes expressed during early autogamy, it is tempting to speculate that these IESs may play
638 a role in controlling the IES excision machinery itself. Indeed, this machinery must be tightly

639 regulated to ensure that all IESs are efficiently excised, while limiting off-target excision of
640 MAC-destined regions, which occurs occasionally in MAC chromosomes [4,30]. Thus,
641 developmental disruption of genes encoding IES excision factors by the excision machinery
642 may provide a simple regulatory feedback loop to decrease the activity of the IES excision
643 machinery as soon as a large fraction of IESs have been excised: if a given IES drives the
644 expression of a protein factor that is essential for IES excision, then this process is progressively
645 interrupted by the removal of this IES during MAC development. More generally, such IESs
646 may provide an exquisite developmental process to regulate DNA elimination events and /or
647 MAC differentiation.
648 Given that each of these 69 IESs is conserved across species and unique within genomes, it
649 seems *a priori* unlikely that they derive from transposable elements. We propose that these
650 conserved IESs may in fact originate from MAC-destined segments. Indeed, mutations within
651 genes can create sequence motifs recognized by the IES excision machinery, and thereby trigger
652 the deletion of functional elements during MAC development. Such mutations are expected to
653 be generally deleterious, and hence to be counter-selected. But occasionally, they might have
654 been positively selected in some genes because they provide a mean to modulate their
655 expression during MAC development or to generate phenotypic diversity among clones, as
656 previously observed for IESs overlapping mating-type genes in many species [12,13]. This
657 scenario would notably explain the presence of conserved IESs overlapping with cellular genes
658 (e.g. Fig 5B). Thus, we propose that the IES excision machinery, which evolved initially to
659 ensure the efficient removal of selfish genetic elements from the MAC genome, has been
660 exapted during evolution as a new way to regulate the expression of some genes.
661

662 ## IES losses

663 In all species, we observed that the length of IESs is negatively correlated with their age (Fig
664 4, S11 Fig). This pattern is similar to that observed in other eukaryotes, where fixed copies of
665 TEs tend to shrink over time, due to the accumulation of small deletions [39]. For IESs located
666 in non-coding region, this progressive shrinking can ultimately lead to their disappearance.
667 Furthermore, these IESs can also be lost by transformation into a MAC-destined sequence (e.g.
668 via mutations the TA dinucleotides). However, for IESs located within exons, losses can only
669 occur by precise complete deletions that leave the open reading frame intact. We did observe
670 such cases of precise loss (S9 Fig). One possible mechanism is that the IES excision machinery,
671 which is normally at work during MAC development, might occasionally operate within the
672 MIC or zygotic nucleus. An alternative hypothesis is that IESs might be lost from the MIC
673 lineage by gene conversion, through homologous recombination with MAC-derived DNA
674 fragments. Interestingly, this scenario might explain cases where we observed concomitant
675 losses of neighboring IESs (see e.g. IES 5 and 6 in S9 Fig). Further studies will be needed to
676 determine the mechanisms underlying precise IES loss. The rate of IES loss has remained quite
677 stable and relatively low across the phylogeny (Fig 1). Conversely, the rate of IES gains has
678 been much more erratic, characterized by episodic waves of insertions, during which the IES
679 gain rate largely exceeded the loss rate (Fig 1). In the end, the number of IESs reflects the
680 balance between gain and loss rates. Thus, the large number of IESs in *Paramecium* can simply
681 be explained by massive invasions of mobile IESs, followed by periods of lower activity, during
682 which IES copies progressively diverge, and occasionally get lost by deletion from the MIC.
683

684 ## Scenario for the evolution of IESs

685 In most organisms, gene regulatory elements and coding regions constitute a no man's land for
686 TEs, because insertions that disrupt gene function are strongly counter-selected. But in some

687 ciliates, it is possible for mobile elements to proliferate within genes in the MIC genome, as
688 long as they are efficiently and precisely excised during the development of the MAC genome,
689 before genes start to be expressed. DNA transposons encode transposases that allow their
690 mobilization by a 'cut-and-paste' process. Generally, the excision step leaves a few nucleotides
691 at the original insertion site, but one peculiarity of piggyBac transposases is that they can excise
692 copies precisely, without leaving any scar [40]. This feature may have predisposed piggyBac
693 to extend its niche to genic regions in ciliates. We speculate that the very first proto-IESs
694 corresponded to piggyBac elements whose transposase was domesticated to target the
695 developing MAC. As soon as several copies of these proto-IESs have been fixed within genes,
696 then the host organism has become dependent on the activity of the piggyBac transposase to
697 ensure that all these copies are precisely excised from its MAC. This selective pressure would
698 have perfected the domestication of the piggyBac transposase by its host, and then,
699 progressively, the evolution of the other components that contribute to the efficient excision of
700 proto-IESs. Once the IES excision machinery in place in the ancestral *Paramecium* lineage,
701 other families of TEs (including non-autonomous elements) could hijack the machinery and in
702 turn exploit this intragenic niche, eventually creating the tens of thousands of IESs found in
703 present-day *Paramecium* genes. The first steps of this scenario remain speculative, since there
704 are no recognizable traces of piggyBac-related IESs in present-day genomes. But the discovery
705 of thousands of mobile IESs directly demonstrates the major contribution of TEs to the
706 expansion of the IES repertoire.
707
708 The coexistence of MAC and MIC is a common feature of all ciliates, yet they do not all contain
709 such a high density of IESs in coding regions. Notably, there are ~12,000 IESs in the germline
710 genome of *Tetrahymena thermophila* (~0.1 IES per kb of MAC-destined sequence), but only
711 11 of them are located within coding regions [41]. These exonic IESs differ from other IESs by

712 their strongly conserved terminal inverted repeats ending with 5'-TTAA-3', the target site of
713 piggyBac transposons. They are excised precisely (restoring a single TTAA) by two
714 domesticated piggyBac transposases, Tpb1 and Tpb6, which may thus have retained the
715 cleavage specificity of their transposon ancestor [42,43]. We analyzed these 11 exonic IESs: 8
716 of them are inserted in protein-coding regions that are not conserved in *Paramecium*, and the
717 other 3 are inserted at sites that do not contain IESs in *Paramecium*. There is therefore no
718 evidence for shared exonic IESs between *T. thermophila* and *Paramecium*. The vast majority
719 of the ~12,000 *T. thermophila* IESs are excised by another domesticated piggyBac transposase,
720 Tpb2 [44]. Although Tpb2 retains the cleavage geometry of piggyBac transposases, producing
721 staggered double-strand breaks with 4-nt 5' overhangs [44], it has lost almost all sequence
722 specificity and is thought to be recruited at IES ends by chromatin marks [45]. As a result,
723 several possible cleavage sites are usually present at IES ends and the rejoining of flanking
724 sequences generates microheterogeneity in the MAC sequence [41], which explains why Tpb2-
725 dependent IESs are restricted to introns and intergenic regions [41]. It is important to note that
726 Tpb2 is an essential gene in *T. thermophila* [44], suggesting that genome-wide retention of IESs
727 in the MAC is still highly detrimental. Interestingly, phylogenetic analyses indicate that the
728 *Paramecium* endonuclease PiggyMac (Pgm) and Tpb2 are more closely related to each other
729 than to Tpb1 or Tpb6, and may even be orthologs [7]. In the case of Pgm, however, sequence
730 specificity was relaxed only for the two distal positions of the 4-nt cleavage sites, and the central
731 TAs remain a strict requirement for IES excision in *Paramecium*. Although piggyBac
732 transposons are completely absent from the present-day *Paramecium* germline, this
733 evolutionary solution may have been favored because it also allowed for precise excision of
734 Tc1/mariner insertions, which in turn would have allowed continuous accumulation of
735 insertions within exons [4].
736

737 Importantly, the fact that a mechanism of precise excision exists in *T. thermophila* (via Tpb1
738 and Tpb6) raises the question of why intragenic IESs are not more abundant in its genome. A
739 similar question arises from the distribution of introns in eukaryotes: why are introns very
740 abundant in some lineages but rare in others (e.g. ~7 introns per gene in vertebrates vs ~0.04 in
741 hemiascomycetous yeast)? Part of the explanation may reside in the fact that, because of
742 population genetic forces, some lineages are more subject to random genetic drift than others,
743 and therefore are more permissive to invasion by weakly deleterious genetic elements [46,47].
744 It is also possible that the abundance of intragenomic parasites is strongly affected by
745 contingency – rare events of massive invasion, followed by long periods during which copies
746 are lost at a slow rate.
747

748 ## Parallel scenario for the evolution of IESs and spliceosomal introns

749 The above scenario for the origin of IESs in *Paramecium* presents an interesting parallel with
750 the one proposed for the evolution of spliceosomal introns in eukaryotes. Indeed, it had long
751 been postulated, based on similarities in biochemical processes, that spliceosomal introns
752 derive from mobile elements (group II self-splicing introns) [48]. In eukaryotes, the spread of
753 introns in protein-coding genes has been facilitated by the fact that transcription and translation
754 occur in separate compartments, thus offering the opportunity for these mobile elements to be
755 excised from the mRNA in the nucleus without interfering with its translation in the cytoplasm
756 [48] – like IESs, which are excised from genes before they get expressed in the MAC. Once the
757 first introns were established, selection drove the emergence of host factors contributing to the
758 efficiency of the splicing process, which progressively led to the evolution of the modern
759 spliceosome [49]. In turn, the existence of the spliceosome released the requirement for introns
760 to maintain their self-splicing activity [50], and allowed other TEs to hijack this machinery.
761 The recent discovery of non-autonomous DNA transposons that generated thousands of introns

762 in genomes of algae directly demonstrated that mobile elements are a major source of new
763 introns [51]. Like IESs, introns represent a burden for their host, because of errors of the
764 splicing machinery [46,47,52–54], but they also contributed to innovations. Indeed, the
765 spliceosome has been exapted during evolution to produce alternative splicing, which not only
766 contributed to diversify the protein repertoire [55] but also allowed new modes of post-
767 transcriptional regulation gene expression [56], in particular in genes that encode splicing
768 factors [57,58]. This pattern is reminiscent of highly conserved IESs that we uncovered in
769 *Paramecium lineages*, which appear to be particularly enriched within genes that are expressed
770 during early MAC development.
771
772 In summary, the evolution of the nuclear envelope opened the way for introns to invade genes
773 in eukaryotes [48], and likewise, the separation of somatic and germline functions between the
774 MIC and the MAC offered the possibility for selfish genetic elements to invade genes in ciliates.
775 Genetic conflicts between these selfish elements and their host genome resulted in the evolution
776 of complex cellular machineries (the spliceosome, the IES excision machinery), which, in the
777 short term, reduced excision errors, but in the long term facilitated proliferation of these
778 elements within genes. The paradigm of intragenomic parasites [59–61] provides a simple and
779 powerful explanation for the "raison d'être" of these mysterious pieces of non-coding DNA
780 that interrupt genes.
781

782 ## Materials and Methods

783 ### Cells and cultivation

784 All experiments were carried out with the *Paramecium* strains listed in Table 1. *Paramecium*
785 *aurelia* cells were grown in a wheat grass powder (WGP, Pines International, USA) infusion
786 medium bacterized the day before use with *Klebsiella pneumoniae* and supplemented with 0.8
787 mg/L of β-sitosterol (Merck). Cultivation and autogamy were carried out at 27 °C. Monoclonal
788 cultures of the *P. caudatum* cells were grown in a 0.25% Cerophyl infusion inoculated with
789 *Enterobacter aerogenes* at 22°C [62].

790

791 ### Micronucleus-enriched preparation

792 To purify the MICs from vegetative cells, we used the same strategy as the one previously
793 published [5,19], with some optimization for the sorting steps. For *Paramecium aurelia*,
794 transgenic cells expressing a micronuclear (MIC)-localized version of the Green Fluorescent
795 Protein (GFP) were obtained by microinjection of the vegetative macronucleus with the *P.*
796 *tetraurelia* CenH3a-GFP plasmid, described in [63]. In the transformed clones, GFP was
797 exclusively found in the MICs and the transformed clones were selected for their GFP
798 signal/noise ratio. Viability of the sexual progeny after autogamy of the transformed clones was
799 systematically monitored to make sure that the presence of the transgene did not impair the
800 functionality of the MICs. A MIC-enriched preparation was obtained from approximately 3 L
801 of exponentially growing vegetative cells after fractionation and Percoll density gradient
802 centrifugation as described in [5] and kept at -80°C until further use.
803 A slightly different procedure was used for *Paramecium caudatum* cells, which were not
804 transformed with the CenH3a-GFP transgene. The MICs of *P. caudatum* strain My43c3d (used
805 for genome sequencing) were purified with a protocol modified from [64]. Briefly, 3L of a
806 starved culture (~600 cells/mL) were filtered through 8 layers of gauze and concentrated by

807 centrifugation in pear-shaped centrifuge tubes. Packed cells were transferred to a 250 mL cell
808 culture flask, resuspended in 150 mL sterile Eau de Volvic and incubated over night at 22°C.
809 All subsequent steps were performed at 4°C or on ice. The overnight culture was again
810 concentrated by centrifugation and the cell pellet was resuspended and washed in 0.25 M TSCM
811 buffer (10 mM Tris-HCl, pH 6.8, 0.25 M sucrose, 3 mM CaCl2, 8mM MgCl2) [65]. After
812 centrifugation for 3 min at 100 g, pelleted cells were resuspended and incubated for 5 min in
813 10 mL 0.25M sucrose-lysis buffer (10 mM Tris-HCl, pH 6.8, 0.25 M sucrose, 3 mM CaCl2,
814 1mM MgCl2, 0.1% Nonidet-P40, 0.1% Na-deoxycholate). The cell suspension was centrifuged
815 for 2 min at 500 g and the packed cells were lysed in 1 mL of 0.25M sucrose-lysis buffer by
816 about 10-20 strokes on a vortex machine. Lysed cells were washed in 14 mL of 0.25 M TSCM
817 buffer and centrifuged for 1 min at 100 g. The supernatant (containing the MICs) was
818 centrifuged for 10 min at 1,500 g and the pellet was resuspended in 8 mL of 60% Percoll. This
819 suspension was centrifuged for 15 min at 24,000 g in a fixed-angle rotor and the micronuclei
820 formed a diffuse band near the middle of the centrifuge tube. This MIC containing layer was
821 carefully removed with a pipette in about 2 mL, diluted with 10 mL of 0.25 M TSCM buffer
822 and pelleted by centrifugation for 10 min at 1,500 g. The MIC pellet was resuspended in 100
823 µL of 0.25 M TSCM buffer, carefully mixed with 50 µL of 50% glycerol and kept at -80°C
824 until further use.
825 The MICs of the other *P. caudatum* strains were purified with a similar protocol, but omitting
826 the Percoll step and replacing it with centrifugation across a sucrose cushion. Lysed cells were
827 resuspended in 9 mL of 0.25 M TSCM buffer and this suspension was carefully layered on top
828 of a sucrose cushion consisting of 2 mL of 1.6 M TSCM buffer and 2 mL of 0.9 M TSCM
829 buffer and centrifuged in a swinging bucket rotor for 10 min at 300 g with lowest acceleration
830 and braking levels. Depending on the strain, the micronuclei accumulated at the bottom of the
831 0.25 M or 0.9 M TSCM cushion and were removed by careful pipetting of the respective phases

832 to new 15-mL tubes. MIC- containing suspensions were diluted with 0.25 M TSCM buffer,
833 centrifuged for 10 min at 1,500 g and the MIC pellets were subsequently treated as described
834 above.
835

836 ## Quantification of MIC DNA content by flow cytometry

837 We measured the absolute DNA content in the nuclei with propidium iodide, a fluorophore that
838 is insensitive to differences in base composition, and compared DNA content of MIC-enriched
839 preparations to a standard (tomato nuclei) of known genome size (see S1 Data). MIC-enriched
840 samples were thawed on ice, diluted 1/5 to 1/10 in washing buffer (0.25 M sucrose; 10 mM Tris
841 pH 7.4; 5 mM MgCl2; 15 mM NaCl; 60 mM KCl; 0.5 mM EGTA) and stained on ice with
842 propidium iodide at 100 µg/mL final concentration. We used Tomato nuclei obtained from
843 Montfavet 63-5 hybrid F1 seeds as internal standards of known genome size. Tomato nuclei
844 were obtained from 1 cm^2 of young leaves chopped in a Petri dish with a scalpel. 800 µL of a
845 modified Galbraith buffer [66], containing 45 mM MgCl2, 30 mM Sodium-Citrate and 20 mM
846 MOPS pH 7.0, 40 µg/mL RNAse A, 0.1% Triton X-100, 5 mM sodium metabisulfite ($S_2O_5Na_2$)
847 was added. The nuclei were collected by pipetting, filtered on 70 µm mesh, and stained on ice
848 with propidium iodide at 100 µg/mL final concentration.
849 The samples were analyzed on a CyanADP Cytomation analyzer from Beckman-Coulter
850 equipped with 3 lasers: 405 nm, 488 nm and 635 nm. Fluorescence intensity (PE signal in pulse-
851 height) of the nuclei was measured at 575/25 nm, after excitation with the 488 nm laser. Results
852 are deduced from 2C nuclei in individuals considered diploid and are given as C-values [67].
853 The ratio of fluorescence intensity of 2C-nuclei from sample and standard allows calculation
854 of genome size. C corresponds to the nuclear genome size (the whole chromosome complement
855 with chromosome number n), 1C and 2C being, respectively, the DNA contents of the haploid
856 (n) and diploid (2n) sets of chromosomes. The haploid nuclear DNA content is expressed in

857 picograms or million base pairs, where 1 pg = 978 Mbp [68], considering Tomato 2C DNA
858 (pg) =1.99, according to [69]. The raw data and calculations are provided in S1 Data.
859

860 ## Micronucleus sorting by flow cytometry and flow imaging

861 To sort the MICs, the MIC-enriched samples were submitted to flow cytometry. Sorting by
862 flow cytometry allowed the separation of the small, diploid MICs from the highly polyploid
863 MAC and the bacteria abundant in *Paramecium* cultures. *P. aurelia* MICs were sorted based
864 on the SSC, FSC, DAPI (DNA staining), and GFP signals. *P. caudatum* MICs, which are bigger
865 than *aurelia* MICs, could be sorted based on their SSC, FSC, and DAPI signals, without the
866 use of a MIC-specific GFP fluorophore. Quality control was performed by flow cell imaging,
867 using the ImageStreamX (Amnis/Merck Millipore) imaging flow cytometer, as previously
868 described [5]. The MICs represented >99% of the sorted sample, except for *P. sonneborni*
869 (97%). An example of sorting is shown in Fig S1.
870

871 ## Genomic DNA extraction and sequencing

872 For MAC DNA sequencing, genomic DNA was extracted from vegetative *Paramecium* cell
873 culture after centrifugation and washes with Tris 10mM pH 7.4. For MIC DNA sequencing,
874 DNA was extracted from the sorted MIC samples. The cell or nuclei pellet was treated with 3
875 volumes of proteinase K solution (0.5 M EDTA pH 9; 1% N-lauroylsarcosine; 1% SDS; 1
876 mg/mL proteinase K) at 55 °C overnight. Genomic DNA was extracted with Tris–HCl-phenol
877 pH 8 with gentle agitation followed by dialysis against TE (10 mM Tris–HCl; 1 mM EDTA,
878 pH 8) 25% ethanol then against Tris 1 mM pH 8. An RNAse A treatment was performed on
879 MAC DNA, followed by phenol extraction and dialysis as described above. DNA concentration
880 was quantified using QuBit High sensibility kit (Invitrogen) and stored at 4 °C.

881 As the amounts of DNA extracted from the MIC are too low (30-50 ng), only an overlapping
882 paired-end library could be prepared for *de novo* sequencing. Briefly, 30-50 ng of MIC DNA
883 were sonicated using the E210 Covaris instrument (Covaris, Inc., USA) in order to generate
884 fragments mostly around 500bp. Illumina libraries were then prepared using the NEBNext
885 DNA Sample Prep Master Mix Set (New England Biolabs, MA, USA) and DNA fragments
886 were PCR-amplified using Platinum Pfx DNA polymerase (Invitrogen) and P5 and P7 primers.
887 Amplified library fragments of roughly 500 – 600 bp were size selected on 2% agarose gel.
888 Libraries traces were validated on a Agilent 2100 Bioanalyzer (Agilent Technologies, USA)
889 and quantified by qPCR using the KAPA Library Quantification Kit (KapaBiosystems) on a
890 MxPro instrument (Agilent Technologies, USA). The libraries were sequenced using 251 base-
891 length read chemistry in a paired-end flow cell on the Illumina HiSeq2500 sequencer (Illumina,
892 USA).
893 For the MAC genomes, an overlapping paired-end library as described above and four
894 additional mate-pair libraries (about 5 kb, 8 kb, 11 kb and 13 kb) were prepared following
895 Nextera protocol (Nextera Mate Pair sample preparation kit, Illumina). Each library was
896 sequenced using 100 base-length read chemistry on a paired-end flow cell on the Illumina
897 HiSeq2000 (Illumina, USA).
898 Information about the sequencing data generated for this study is available in S7 Table.
899

900 ### RNA extraction and sequencing

901 For the purpose of gene annotation, we sequenced mRNAs from vegetative cells (S7 Table).
902 400 mL cultures of exponentially growing cells at 1000 cells/mL were centrifuged and flash-
903 frozen in liquid nitrogen prior to TRIzol (Invitrogen) treatment, modified by the addition of
904 glass beads for the initial lysis step.

905 RNA-Seq library preparation was carried out from 1 µg total RNA using the TruSeq Stranded
906 mRNA kit (Illumina, San Diego, CA, USA), which allows mRNA strand orientation (sequence
907 reads occur in the same orientation as anti-sense RNA). Briefly, poly(A)+ RNA was selected
908 with oligo(dT) beads, chemically fragmented and converted into single-stranded cDNA using
909 random hexamer priming. Then, the second strand was generated to create double-stranded
910 cDNA. cDNAs were then 3'-adenylated, and Illumina adapters were added. Ligation products
911 were PCR-amplified. Ready-to-sequence Illumina libraries were then quantified by qPCR using
912 the KAPA Library Quantification Kit for Illumina Libraries (KapaBiosystems, Wilmington,
913 MA, USA), and library profiles evaluated with an Agilent 2100 Bioanalyzer (Agilent
914 Technologies, Santa Clara, CA, USA). Each library was sequenced using 101 bp paired end
915 read chemistry on a HiSeq2000 Illumina sequencer.
916

917 ### MAC Genome assembly

918 The MAC genomes sequenced for this project were all assembled according to the following
919 steps.
920 First, long Illumina reads were obtained from 250 bp overlapping paired-end reads sequenced
921 from ~450 bp fragments. The reads were fused with fastx_mergepairs, an in-house tool
922 developed at Genoscope using the fastx library (http://hannonlab.cshl.edu/fastx_toolkit/). An
923 alignment of at least 15 bp with at least 90% identity and fewer than 4 errors was required to
924 fuse two reads into one longer read. The set of fused reads, completed with any reads that could
925 not be fused, was assembled into contigs by the Newbler version 2.9 overlap-layout-consensus
926 assembler, with a minimal alignment identity of 99% and a minimal alignment size of 99 bp.
927 Scaffolds were built from the contigs using 4 Illumina mate-pair libraries with respective insert
928 sizes of 5 kb, 8 kb, 11 kb and 13 kb. The scaffolder SSpace [70] was used, with default
929 parameters and an acceptable variation in mate-pair insert size of 25%. Gap closing was a two

930 step process with SOAPdenovo2 GapCloser software [71]. The first step used the Illumina
931 paired-end reads, the second step used the Illumina mate-pair libraries. Finally, Kraken software
932 [72] and the NR nucleotide database were used to detect and remove non-eukaryotic scaffolds,
933 owing mainly to bacterial contaminants (see below).
934

935 ### Filtering

936 Scaffolds with a length inferior to 2 kb or with a G+C content greater than 40% were filtered.
937 Contaminant scaffolds were identified and removed from the assembly provided the Kraken
938 Kmer score was superior to 10 or a BLASTN match (-evalue 1e-40 –perc_identity 70) against
939 RefSeq database (excluding *Paramecium* sequences) covered at least 20% of the scaffold
940 length. If the mitochondrial genome (more or less fragmented) could be identified by a
941 BLASTN (-evalue 1e-1 –perc_identity 70) against the *P. tetraurelia* mitochondrial genome, the
942 scaffold(s) were tagged as mitochondrial. A handful of chimeric scaffolds were detected and
943 corrected in the *P. octaurelia*, *P. primaurelia* and *P. sexaurelia* assemblies by visual inspection
944 of available long-range sequencing data (remapped mate-pairs) (see S2 Data)).
945

946 ### The constitutive MAC

947 Paired-end MAC DNA sequencing data were mapped on the MAC genome assembly using
948 Bowtie2 (v.2.2.3 –local, otherwise default parameters) [73]. We defined the constitutive MAC
949 as consisting of all regions of the assemblies with the expected average read coverage.We
950 defined the regions of low coverage at scaffold extremities as MAC-variable regions. These
951 regions result from the variability of programmed genome rearrangement patterns during MAC
952 development [74]. While most MIC loci are either fully eliminated during MAC development
953 (MIC-limited sequences) or fully retained (MAC-destined sequences), MAC-variable regions
954 correspond to DNA sequences that are not completely eliminated and instead, are retained in a

955 small fraction of MAC copies. MAC-variable regions represent ~15% of the initial MAC
956 genome assembly (S2 Table). In relation to the MAC DNA-seq depth, a minimum expected
957 coverage (v1.9 of samtools depth –q 10 –Q 10) was defined for each assembly (*P. octaurelia*
958 50X, *P. pentaurelia* 35X, *P. primaurelia* 20X, and *P. sonneborni* 35X). For each scaffold
959 extremity, a Perl script analyzed the coverage in sliding 2 kb windows. The first window from
960 the end of the scaffold with a coverage above the minimum expected coverage delimitated the
961 end of the MAC-variable regions. Only regions of minimum size 4 kb were kept. The script
962 adjusted region ends using the MAC telomerisation sites and the ends of coding genes. After
963 this automatic pipeline, each scaffold and mask was adjusted by eye using Circos drawings [75]
964 (see example S6 Fig, representing DNA and RNA coverage, density in non-coding genes and
965 positions of the MAC telomerisation sites). The positions of the regions used to reconstruct the
966 constitutive MAC for each MAC assembly are provided in S2 Data.
967

968 ### IES annotation

969 Annotation of IESs was performed using the ParTIES toolkit [21] with default parameters.
970 Briefly, this involves (i) alignment of MIC paired-end reads with a reference MAC genome to
971 establish a catalog of potential IES insertion sites and to exclude reads that match perfectly
972 across these sites hence do not contain IESs; (ii) assembly of the remaining reads with Velvet
973 to obtain contigs that may contain IESs; (iii) alignment of the contigs with the MAC reference
974 genome to determine the position and the sequence of the IESs.
975

976 ### Gene annotation

977 Gene annotation for the 9 species was carried out using a pipeline specifically tuned for the
978 high gene density and tiny intron size (20 – 30 nt) characteristic of *Paramecium* somatic
979 genomes. RNA-Seq transcriptome data was used to predict transcription units with the TrUC

980 v1.0 software (https://github.com/oarnaiz/TrUC), as detailed in [37]. EuGene v4.1 software
981 [76] configured with curated *Paramecium tetraurelia* genes [37] was used for *ab initio*
982 predictions and to combine annotation evidence (the transcription units, the *ab initio* predictions
983 and comparative genomics evidence). Gene annotation completeness was calculated using
984 BUSCO (v4.1.4 --mode prot --lineage_dataset alveolate; [77]) through the Galaxy Portal
985 (https://usegalaxy.org/).
986

987 ## Assembly-free Genome Size Estimation

988 Illumina paired-end sequencing reads were used to estimate genome size based on counting all
989 substrings of 17 nt in the reads, using jellyfish software version 2.2.10 [78]:
990

991 ```
jellyfish count -t 12 -C -m 17 -s 5G -o <sample.jf> <sample_paired_end_reads.fastq>
```
992 ```
jellyfish histo -o <sample.histo> <sample.jf>
```
993

994 The method for genome size estimation, described in [18,19], assumes that the total number
995 of *k*-mers (in this case 17-mers) divided by the sequencing depth is a good approximation for
996 genome size.
997 As discussed in [19] for *Paramecium* genomes, the histogram of *k*-mer depth for a perfect,
998 homozygous genome with no repeated sequences (and no sequencing errors) is fit by a Poisson
999 distribution, the peak corresponding to sequencing depth. The estimate of genome size is
1000 obtained by dividing the total *k*-mer count (excluding the peak near the origin that results from
1001 *k*-mers with sequencing errors) by the sequencing depth. This is straightforward for MAC
1002 genomes. As shown in S2 Table, the estimated MAC genome sizes were in good agreement
1003 with the size of the constitutive MAC genome assemblies.
1004 For MIC genomes, variable amounts of contamination from MAC DNA lead to a second peak
1005 at higher *k*-mer depth corresponding to the sum of MAC-destined k-mers in the MIC DNA and

1006 the MAC *k*-mers in the contaminating MAC DNA. This was only a significant problem for the
1007 *P. tetraurelia*, *P. sexaurelia* and *P. sonneborni* MIC DNA samples, which were approximately
1008 corrected by assuming that a proportion of the *k*-mers counted from this second peak up to a
1009 depth of 500 were contributed by the contaminating MAC reads, while all the *k*-mers with a
1010 depth greater than 500, corresponding to highly repeated sequences, are of MIC origin (S2 Fig).
1011 The proportion of contaminating MAC DNA needed for this calculation was confirmed using
1012 IES retention scores (IRS) calculated with the MIC sequencing reads [21]. The position of the
1013 peak in the IRS distribution indicates the proportion of MIC (IRS ~ 1) and MAC (IRS ~ 0)
1014 DNA in the sample, as illustrated in S2 Fig.
1015

1016 ### Identification of gene families

1017 We performed an all against all BLASTP (ncbi-blast+ v. 2.2.30+) [79] search using the
1018 predicted protein sequences from each genome including also the proteins of *Tetrahymena*
1019 *thermophila* (June 2014 assembly http://ciliate.org) as an outgroup. From the resulting output
1020 we determined gene families with SiLiX v. 1.2.9 [80]. The resulting gene families were aligned
1021 with MAFFT v7.305b (2016/Aug/16) [81] using the --auto option. Gene families with fewer
1022 than 3 genes or average pairwise identity less than 50% were excluded from downstream
1023 analyses. From the protein alignments we reconstructed the nucleotide coding sequence
1024 alignments.
1025

1026 ### Paramecium species phylogeny

1027 To reconstruct the species phylogeny, we first selected single-copy gene families present in all
1028 nine *Paramecium* species (N=1,061 genes). When available, the *T. thermophila* homolog was
1029 also included as an outgroup. We estimated the maximum likelihood phylogeny using IQtree
1030 v.1.4.2 [82], considering each gene as a separate partition. We performed model testing on each

1031 partition and chose the best codon model (determined by the largest BIC). We evaluated the
1032 results by 1,000 bootstrap replicates. All internal branches but one are supported by 100%
1033 bootstrap values (Fig 1). We will hereafter refer to this species tree inferred from single-copy
1034 gene families as *Tree1*.
1035
1036 The rationale for analyzing single-copy gene families is that these sets of homologous
1037 sequences are *a priori* expected to correspond to orthologs. However, given that *Paramecium*
1038 have been subject to three rounds of whole genome duplications followed by massive gene
1039 losses [83], it is possible that some single-copy gene families include paralogs. To check
1040 whether hidden paralogs might have biased the estimation of the species tree, we used
1041 PHYLDOG v.2.0beta (build 10/10/2016), a maximum likelihood method to jointly infer rooted
1042 species and gene trees, accounting for gene duplications and losses [25]. The analysis was
1043 performed using all gene families (N=13,617). The default program options were used with
1044 additionally setting a random starting species tree and BIONJ starting gene trees. The
1045 duplication and loss parameters were optimized with the average then branchwise option and
1046 the genomes were not assumed to have the same number of genes. We also ran PHYLDOG
1047 considering *Tree1* as the fixed species tree, and keeping the remaining options identical. The
1048 topology of the most likely species tree inferred with PHYLDOG is almost identical to *Tree1*
1049 (it only slightly differs in the positions of *P. biaurelia* and *P. tredecaurelia*), and its likelihood
1050 is not significantly different from that obtained when running PHYLDOG with *Tree1* as a
1051 species tree. Thus, the species tree inferred by PHYLDOG using all gene families (N=13,617)
1052 shows no significant disagreement with the phylogeny based on single-copy gene families
1053 (*Tree1*). We therefore hereafter considered *Tree1* as the reference species tree for all our
1054 analyses. To identify duplication and speciation nodes in gene phylogenies, we computed
1055 reconciled trees for each gene family with PHYLDOG, using *Tree1* as a species tree.

1056

1057

1058 ### Taking into account the uncertainty of IES presence due to limited detection sensitivity.

1059 To identify events of IES gain and loss along the species phylogeny, it is necessary to analyze
1060 the pattern of presence/absence of IESs at homologous loci across species. One difficulty is that
1061 some IESs may remain undetected (false negatives). In particular, the sensitivity of ParTIEs
1062 depends on the local read coverage [21]. To take into account the uncertainty arising from the
1063 variable local read coverage along scaffolds of each species we calculated the coverage of MIC
1064 reads mapped against the MAC genome. We identified genes with extreme values of coverage
1065 (less than the 10th percentile or more than the 90th over all genes) or with an absolute read
1066 coverage of less than 15 reads. These genes correspond to regions with possible assembly errors
1067 or to regions of low power to detect IESs, and we marked them as problematic for IES
1068 annotation. IESs in these genes were considered to have an uncertain status of presence, and if
1069 no IES was annotated the genes were marked as potentially containing IESs. To avoid issues
1070 due to genome assembly errors, we excluded from our analyses all IESs identified on small
1071 scaffolds (< 10 kb).

1072

1073 ### Taking into account the uncertainty of IES location (floating IESs).

1074 To identify co-orthologous IES loci, i.e. that result from a single ancestral insertion event, we
1075 searched for IESs located at a same site across homologous sequences. It should be noted that
1076 the location of IESs, inferred from the comparison of MIC and MAC sequences, is sometimes
1077 ambiguous. This occurs when the IES boundaries overlap a motif repeated in tandem (S13 Fig).
1078 Such cases, hereafter called "floating IESs", represent 7% of all IESs. In the vast majority of
1079 cases (86%) the alternative locations of floating IESs differ by only two bp (as in the example
1080 shown in S13 Fig), and there are less than 1% of floating IESs for which the uncertainty in IES

1081 position exceeds 5 bp. To determine the exact location of IESs and capture the inherent
1082 ambiguity due to possible floating IESs we used a 10 bp window around each annotated IES
1083 location to determine if the IES was classified as floating. If so, the alternative locations were
1084 added to the IES annotation.
1085

1086 ### Co-orthologous IES insertion sites

1087 To detect co-orthologous IES loci, we compared the position of IESs within homologous genes.
1088 To do so, we analyzed gene families with more than 3 sequences and average pairwise identity
1089 (at the protein sequence level) of more than 50%. To avoid ambiguity in the identification of
1090 homologous sites, we filtered protein sequence alignments with GBlocks v0.91b [84] and we
1091 only retained IESs located within conserved alignment blocks. An IES insertion site spans two
1092 nucleotides (TA). In a multiple sequence alignment including gaps, an IES locus can be larger
1093 (e.g. T--A). Two IES loci were considered as homologous if they have at least one shared site
1094 within the alignment (taking into account all potential locations in the case of floating IESs). In
1095 the case of floating IESs overlapping the boundaries of conserved alignment blocks, the
1096 presence or absence of homologous IES loci in other sequences cannot be reliably inferred. We
1097 therefore only retained IESs for which all homologs (if any) are entirely located within the
1098 conserved alignment blocks (i.e. we discarded sets of co-orthologous IES loci that included
1099 some floating IESs for which some of the possible alternative positions were located outside of
1100 the conserved alignment blocks).
1101

1102 ### Ancestral state reconstruction and inference of IES insertion and loss rates

1103 To explore the dynamics of IES gain and loss we used a Bayesian approach to reconstruct the
1104 ancestral states of presence and absence of IESs using revBayes 1.0.0 beta 3 (2015-10-02) [85].
1105 We constructed binary character matrices (presence/absence) for each gene family containing

1106 at least one IES unambiguously located within a conserved alignment block (see above). We
1107 assumed a model of character evolution with one rate of gain and one rate of loss sampled from
1108 the same exponential distribution with parameter α and a hyperprior sampled from an
1109 exponential with parameter 1. We excluded from the analysis 5 gene families for which
1110 revBayes could not compute a starting probability due to very small numbers. We used
1111 PHYLDOG reconciled gene trees (see above) to fix gene tree topologies and branch lengths.
1112 We ran $5x10^5$ iterations. The search parameters were optimized in an initial phase of 10,000
1113 iterations with tuning interval 1,000. Good sampling of the parameter space was verified by
1114 inspecting the time series and autocorrelation plots of the parameters. The convergence was
1115 validated by inspecting the multivariate Gelman and Rubin's diagnostic plots for different
1116 iterations.
1117
1118 Thus, for a given IES locus in a given gene family, revBayes provides an estimate of the
1119 probability of presence of an IES at each node of the gene phylogeny. We used these
1120 probabilities of presence along the gene phylogeny to estimate rates of IES gains or losses in
1121 each branch of the species tree. Because of gene duplications, a given branch in the species tree
1122 can be represented by several paths in the gene tree. Thus, we considered all paths in the gene
1123 tree that connect the corresponding speciation nodes (see S8 Fig for a simplified example). To
1124 measure the IES gain rate at a given IES locus (*c*), in a given gene family (*g*), we define p^+_{cgij}
1125 as the sum of increase in probability of presence of an IES at this locus along all paths of gene
1126 family *g* connecting speciation nodes *i* and *j* (where *i* is a direct ancestor of *j*). Let *ng* be the
1127 length in kilobase pairs of gene family *g* alignment (counting only well aligned sites, where the
1128 presence of IESs can be assessed). Let *Ig* be the number of IES loci in family *g*. Let *kgij* be the
1129 number of paths connecting speciation nodes *i* and *j* in family *g*. Let *bij* be the branch length
1130 connecting nodes *i* and *j* in the species tree (*bij* is taken here as a proxy for time). Let p^+_{gij} be

1131 the sum of increase in probability of presence of an IES, cumulated over all IES loci in family
1132 g. We define p^{+}_{ij} as the sum of increase in probability of presence of an IES, cumulated over
1133 all IES loci along the path i to j of family g. We define Gij as the rate of IES gain over all gene
1134 families (f) along path ij expressed in number of IES gains per kilobase pairs of alignment per
1135 unit of time:

1136
$$G_{ij} = \frac{\sum_{g=1}^{g=f} p^{+}_{gij}}{\sum_{g=1}^{g=f} n_g k_{gij} \cdot b_{ij}}$$

1137 We define in a similar manner the rate of IES loss. For a given gene family g, let p^{-}_{cgij} be the
1138 sum of decreases in probability of presence of an IES in IES locus c along a lineage in gene
1139 family g connecting speciation nodes i and j. Let Ig be the number of IES loci in family g. Let
1140 p^{-}_{gij} be the sum of decrease in probability of presence of an IES, cumulated over all IES loci in
1141 family g. We define as Lij the rate of IES loss over all gene families (f) along path ij expressed
1142 in number of IES losses per IES, per unit of time.
1143

1144
$$L_{ij} = \frac{\sum_{g=1}^{g=f} p^{-}_{g_{ij}}}{\sum_{g=1}^{g=f} I_g \cdot kg_{ij} \cdot b_{ij}}$$

1145

1146 ## IES age of insertion

1147 The age of first insertion for each group of co-orthologous IES locations is defined as the age
1148 of the most recent common ancestor of all nodes in which an ancestral IES was present with
1149 probability larger than 99%.
1150

1151 ## Identification of homologous IES sequences and characterization of mobile IESs

1152 To characterize families of homologous IES sequences, we first compared all IESs (from all
1153 species) against each other with *blastn* (ncbi blast+ v2.5.0, [79]):

```
blastn -evalue 1e-8 -query IES.fa -db IES -dust yes -task blastn -
max_target_seqs 10000
```

We retained all pairs of homologous IESs for which BLAST alignments encompass the first and last 20 nt of the query and subject sequences. This ensures that the detected sequence homology includes the boundaries of the IESs, and is not merely due to the presence of repeated sequences inserted within a pre-existing IES.

To identify potentially mobile IESs, we searched for homologous IES sequences present at different (non-homologous) genomic loci. For this, we extracted 100 nt on each side of the IES location, and compared all these flanking regions against each other with *blastn* (using the same parameters as above). Pairs of homologous IES sequences with strong hits in flanking regions (≥75% identity over 150 nt or more) were classified as 'homologous IESs at homologous loci'. The other pairs were classified as 'candidate mobile IESs'. We clustered each group based on sequence similarity using SiLiX [80] with default parameters.

We further analyzed all clusters of candidate mobile IESs having at least 10 sequences (N=57 clusters). For each cluster, we constructed multiple sequence alignments with MAFFT v7.305b (with `--adjustdirection` and `--auto` options). We manually inspected these alignments to select full-length copies and create a multiple alignment covering the entire repeated element. At this stage, we excluded 11 clusters corresponding to very AT-rich sequences, for which it was not clear whether the detected sequence similarities were due to homology or to their highly biased sequence composition. Furthermore, two clusters were split into subfamilies, to include only sequences that are homologous over their entire length. We

then used these seed alignments to build an HMM profile for each repeat family and search for homologous copies among the entire IES dataset with NHMMER version 3.1b2 [86].

In total, NHMMER identified 12,184 IESs having a significant hit (E-value < 10^{-3}) in the dataset of HMM profiles. Among detected hits, we distinguished two categories: 1) cases where the detected repeated element is located within the IES but does not overlap with the extremities of the IES (i.e. nested repeats), and 2) cases where the extremities of the HMM profile align with the extremities of the IES (with a tolerance of 3 bp to allow alignment uncertainties). IESs belonging to this latter category were hereafter considered as 'mobile IESs'. For subsequent analyses, we selected all families with more than 10 mobile IESs in at least one genome (N=24 families of mobile IESs). Multiple alignments, HMM profiles and the list of matching IESs are available (https://doi.org/10.5281/zenodo.4726682).

Acknowledgments

We thank Damien de Vienne for his help in phylogenetic analyses, Alexey Potekhin (Saint Petersburg State University, St Petersburg, Russia) and Ewa Przybos (Institute of Systematics and Evolution of Animals, Polish Academy of Sciences, Cracow, Poland) for the maintenance of *Paramecium* stock collections and providing some *Paramecium aurelia* strains. This work was performed using the computing facilities of the CC LBBE/PRABI.

1200 ## References

1201

1202 1. Cheng C-Y, Orias E, Leu J-Y, Turkewitz AP. The evolution of germ-soma nuclear
1203 differentiation in eukaryotic unicells. Curr Biol. 2020;30: R502–R510.
1204 doi:10.1016/j.cub.2020.02.026
1205 2. Betermier M, Duharcourt S. Programmed Rearrangement in Ciliates: Paramecium.
1206 Microbiol Spectr. 2014;2. doi:10.1128/microbiolspec.MDNA3-0035-2014
1207 3. Rzeszutek I, Maurer-Alcalá XX, Nowacki M. Programmed genome rearrangements in
1208 ciliates. Cell Mol Life Sci. 2020. doi:10.1007/s00018-020-03555-2
1209 4. Arnaiz O, Mathy N, Baudry C, Malinsky S, Aury J-M, Denby Wilkes C, et al. The
1210 Paramecium Germline Genome Provides a Niche for Intragenic Parasitic DNA: Evolutionary
1211 Dynamics of Internal Eliminated Sequences. PLoS Genetics. 2012;8: e1002984.
1212 doi:10.1371/journal.pgen.1002984
1213 5. Guérin F, Arnaiz O, Boggetto N, Denby Wilkes C, Meyer E, Sperling L, et al. Flow
1214 cytometry sorting of nuclei enables the first global characterization of Paramecium germline
1215 DNA and transposable elements. BMC Genomics. 2017;18. doi:10.1186/s12864-017-3713-7
1216 6. Baudry C, Malinsky S, Restituito M, Kapusta A, Rosa S, Meyer E, et al. PiggyMac, a
1217 domesticated piggyBac transposase involved in programmed genome rearrangements in the
1218 ciliate Paramecium tetraurelia. Genes Dev. 2009;23: 2478–2483. doi:10.1101/gad.547309
1219 7. Bischerour J, Bhullar S, Denby Wilkes C, Régnier V, Mathy N, Dubois E, et al. Six
1220 domesticated PiggyBac transposases together carry out programmed DNA elimination in
1221 Paramecium. Elife. 2018;7. doi:10.7554/eLife.37927
1222 8. Vanssay A de, Touzeau A, Arnaiz O, Frapporti A, Phipps J, Duharcourt S. The
1223 Paramecium histone chaperone Spt16-1 is required for Pgm endonuclease function in
1224 programmed genome rearrangements. PLOS Genetics. 2020;16: e1008949.

1225 doi:10.1371/journal.pgen.1008949
1226 9. Abello A, Régnier V, Arnaiz O, Le Bars R, Bétermier M, Bischerour J. Functional
1227 diversification of Paramecium Ku80 paralogs safeguards genome integrity during precise
1228 programmed DNA elimination. PLoS Genet. 2020;16: e1008723.
1229 doi:10.1371/journal.pgen.1008723
1230 10. Klobutcher LA, Herrick G. Consensus inverted terminal repeat sequence of
1231 Paramecium IESs: resemblance to termini of Tc1-related and Euplotes Tec transposons.
1232 Nucleic Acids Res. 1995;23: 2006–2013. doi:10.1093/nar/23.11.2006
1233 11. Klobutcher LA, Herrick G. Developmental genome reorganization in ciliated
1234 protozoa: the transposon link. Prog Nucleic Acid Res Mol Biol. 1997;56: 1–62.
1235 doi:10.1016/s0079-6603(08)61001-6
1236 12. Singh DP, Saudemont B, Guglielmi G, Arnaiz O, Goût J-F, Prajer M, et al. Genome-
1237 defence small RNAs exapted for epigenetic mating-type inheritance. Nature. 2014;509: 447–
1238 452. doi:10.1038/nature13318
1239 13. Sawka-Gądek N, Potekhin A, Singh DP, Grevtseva I, Arnaiz O, Penel S, et al.
1240 Evolutionary Plasticity of Mating-Type Determination Mechanisms in Paramecium aurelia
1241 Sibling Species. Genome Biol Evol. 2020;13. doi:10.1093/gbe/evaa258
1242 14. McGrath CL, Gout J-F, Johri P, Doak TG, Lynch M. Differential retention and
1243 divergent resolution of duplicate genes following whole-genome duplication. Genome
1244 Research. 2014;24: 1665–1675. doi:10.1101/gr.173740.114
1245 15. McGrath CL, Gout J-F, Doak TG, Yanagi A, Lynch M. Insights into Three Whole-
1246 Genome Duplications Gleaned from the *Paramecium caudatum* Genome Sequence. Genetics.
1247 2014;197: 1417–1428. doi:10.1534/genetics.114.163287
1248 16. Gout J-F, Johri P, Arnaiz O, Doak TG, Bhullar S, Couloux A, et al. Universal trends
1249 of post-duplication evolution revealed by the genomes of 13 Paramecium species sharing an

1250 ancestral whole-genome duplication. bioRxiv. 2019; 573576. doi:10.1101/573576
1251 17. Sonneborn TM. Paramecium aurelia. In: King RC, editor. Handbook of Genetics:
1252 Plants, Plant Viruses, and Protists. Boston, MA: Springer US; 1974. pp. 469–594.
1253 doi:10.1007/978-1-4684-2994-7_20
1254 18. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and de novo assembly of
1255 the giant panda genome. Nature. 2010;463: 311–317. doi:10.1038/nature08696
1256 19. Duharcourt S, Sperling L. The Challenges of Genome-Wide Studies in a Unicellular
1257 Eukaryote With Two Nuclear Genomes. Meth Enzymol. 2018;612: 101–126.
1258 doi:10.1016/bs.mie.2018.08.012
1259 20. Johri P, Krenek S, Marinov GK, Doak TG, Berendonk TU, Lynch M. Population
1260 Genomics of Paramecium Species. Molecular Biology and Evolution. 2017;34: 1194–1216.
1261 doi:10.1093/molbev/msx074
1262 21. Denby Wilkes C, Arnaiz O, Sperling L. ParTIES: a toolbox for Paramecium
1263 interspersed DNA elimination studies. Bioinformatics. 2016;32: 599–601.
1264 doi:10.1093/bioinformatics/btv691
1265 22. Mayer KM, Mikami K, Forney JD. A mutation in Paramecium tetraurelia reveals
1266 functional and structural features of developmentally excised DNA elements. Genetics.
1267 1998;148: 139–149.
1268 23. Duharcourt S, Keller AM, Meyer E. Homology-dependent maternal inhibition of
1269 developmental excision of internal eliminated sequences in Paramecium tetraurelia. Mol Cell
1270 Biol. 1998;18: 7075–7085. doi:10.1128/mcb.18.12.7075
1271 24. Przyboś E, Tarcz S, Rautian M, Sawka N. Delimiting Species Boundaries within a
1272 Paraphyletic Species Complex: Insights from Morphological, Genetic, and Molecular Data on
1273 Paramecium sonneborni (Paramecium aurelia species complex, Ciliophora, Protozoa). Protist.
1274 2015;166: 438–456. doi:10.1016/j.protis.2015.07.001

1275 25. Boussau B, Szollosi GJ, Duret L, Gouy M, Tannier E, Daubin V. Genome-scale
1276 coestimation of species and gene trees. Genome Research. 2013;23: 323–330.
1277 doi:10.1101/gr.141978.112
1278 26. Przyboś E, Tarcz S. New Stands of the Paramecium aurelia spp. Complex (Protista,
1279 Oligohymenophorea) in Ethiopia, Madagascar, Taiwan, and Romania. Folia Biologica.
1280 2018;66: 111–119. doi:10.3409/fb_66-3.12
1281 27. Przyboś E, Tarcz S, Rautian M, Sawka N. Delimiting Species Boundaries within a
1282 Paraphyletic Species Complex: Insights from Morphological, Genetic, and Molecular Data on
1283 Paramecium sonneborni (Paramecium aurelia species complex, Ciliophora, Protozoa). Protist.
1284 2015;166: 438–456. doi:10.1016/j.protis.2015.07.001
1285 28. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New
1286 algorithms and methods to estimate maximum-likelihood phylogenies: assessing the
1287 performance of PhyML 3.0. Syst Biol. 2010;59: 307–321. doi:10.1093/sysbio/syq010
1288 29. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo
1289 generator. Genome Res. 2004;14: 1188–1190. doi:10.1101/gr.849004
1290 30. Duret L, Cohen J, Jubin C, Dessen P, Gout J-F, Mousset S, et al. Analysis of sequence
1291 variability in the macronuclear DNA of Paramecium tetraurelia: A somatic view of the
1292 germline. Genome Research. 2008;18: 585–596. doi:10.1101/gr.074534.107
1293 31. Lhuillier-Akakpo M, Frapporti A, Denby Wilkes C, Matelot M, Vervoort M, Sperling
1294 L, et al. Local Effect of Enhancer of Zeste-Like Reveals Cooperation of Epigenetic and cis-
1295 Acting Determinants for Zygotic Genome Rearrangements. PLoS Genetics. 2014;10:
1296 e1004665. doi:10.1371/journal.pgen.1004665
1297 32. Maliszewska-Olejniczak K, Gruchota J, Gromadka R, Denby Wilkes C, Arnaiz O,
1298 Mathy N, et al. TFIIS-Dependent Non-coding Transcription Regulates Developmental
1299 Genome Rearrangements. PLoS Genet. 2015;11: e1005383.

1300 doi:10.1371/journal.pgen.1005383
1301 33. Gruchota J, Denby Wilkes C, Arnaiz O, Sperling L, Nowak JK. A meiosis-specific
1302 Spt5 homolog involved in non-coding transcription. Nucleic Acids Res. 2017;45: 4722–4732.
1303 doi:10.1093/nar/gkw1318
1304 34. Sandoval PY, Swart EC, Arambasic M, Nowacki M. Functional diversification of
1305 Dicer-like proteins and small RNAs required for genome sculpting. Dev Cell. 2014;28: 174–
1306 188. doi:10.1016/j.devcel.2013.12.010
1307 35. Frapporti A, Miró Pina C, Arnaiz O, Holoch D, Kawaguchi T, Humbert A, et al. The
1308 Polycomb protein Ezl1 mediates H3K9 and H3K27 methylation to repress transposable
1309 elements in Paramecium. Nature Communications. 2019;10: 2710. doi:10.1038/s41467-019-
1310 10648-5
1311 36. Lepère G, Bétermier M, Meyer E, Duharcourt S. Maternal noncoding transcripts
1312 antagonize the targeting of DNA elimination by scanRNAs in Paramecium tetraurelia. Genes
1313 Dev. 2008;22: 1501–1512. doi:10.1101/gad.473008
1314 37. Arnaiz O, Van Dijk E, Bétermier M, Lhuillier-Akakpo M, de Vanssay A, Duharcourt
1315 S, et al. Improved methods and resources for paramecium genomics: transcription units, gene
1316 annotation and gene expression. BMC Genomics. 2017;18: 483. doi:10.1186/s12864-017-
1317 3887-z
1318 38. Blumenstiel JP. Birth, School, Work, Death, and Resurrection: The Life Stages and
1319 Dynamics of Transposable Element Proliferation. Genes. 2019;10: 336.
1320 doi:10.3390/genes10050336
1321 39. Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. Evidence for DNA Loss as
1322 a Determinant of Genome Size. Science. 2000;287: 1060–1062.
1323 doi:10.1126/science.287.5455.1060
1324 40. Mitra R, Fain-Thornton J, Craig NL. piggyBac can bypass DNA synthesis during cut

1325 and paste transposition. EMBO J. 2008;27: 1097–1109. doi:10.1038/emboj.2008.41
1326 41. Hamilton EP, Kapusta A, Huvos PE, Bidwell SL, Zafar N, Tang H, et al. Structure of
1327 the germline genome of Tetrahymena thermophila and relationship to the massively
1328 rearranged somatic genome. eLife. 2016;5. doi:10.7554/eLife.19090
1329 42. Cheng C-Y, Young JM, Lin C-YG, Chao J-L, Malik HS, Yao M-C. The piggyBac
1330 transposon-derived genes TPB1 and TPB6 mediate essential transposon-like excision during
1331 the developmental rearrangement of key genes in Tetrahymena thermophila. Genes Dev.
1332 2016;30: 2724–2736. doi:10.1101/gad.290460.116
1333 43. Feng L, Wang G, Hamilton EP, Xiong J, Yan G, Chen K, et al. A germline-limited
1334 piggyBac transposase gene is required for precise excision in Tetrahymena genome
1335 rearrangement. Nucleic Acids Research. 2017;45: 9481–9502. doi:10.1093/nar/gkx652
1336 44. Cheng C-Y, Vogt A, Mochizuki K, Yao M-C. A Domesticated piggyBac Transposase
1337 Plays Key Roles in Heterochromatin Dynamics and DNA Cleavage during Programmed DNA
1338 Deletion in Tetrahymena thermophila. MBoC. 2010;21: 1753–1762. doi:10.1091/mbc.e09-
1339 12-1079
1340 45. Vogt A, Mochizuki K. A domesticated PiggyBac transposase interacts with
1341 heterochromatin and catalyzes reproducible DNA elimination in Tetrahymena. PLoS Genet.
1342 2013;9: e1004032. doi:10.1371/journal.pgen.1004032
1343 46. Lynch M. The Origins of Eukaryotic Gene Structure. Molecular Biology and
1344 Evolution. 2006;23: 450–468. doi:10.1093/molbev/msj050
1345 47. Lynch M. Intron evolution as a population-genetic process. Proceedings of the
1346 National Academy of Sciences. 2002;99: 6118–6123. doi:10.1073/pnas.092595699
1347 48. Cavalier-Smith T. Intron phylogeny: a new hypothesis. Trends in Genetics. 1991;7:
1348 145–148. doi:10.1016/0168-9525(91)90377-3
1349 49. Wahl MC, Will CL, Lührmann R. The Spliceosome: Design Principles of a Dynamic

1350 RNP Machine. Cell. 2009;136: 701–718. doi:10.1016/j.cell.2009.02.009
1351 50. Doolittle WF. Is junk DNA bunk? A critique of ENCODE. Proceedings of the
1352 National Academy of Sciences. 2013;110: 5294–5300. doi:10.1073/pnas.1221376110
1353 51. Huff JT, Zilberman D, Roy SW. Mechanism for DNA transposons to generate introns
1354 on genomic scales. Nature. 2016;538: 533–536. doi:10.1038/nature20110
1355 52. Pickrell JK, Pai AA, Gilad Y, Pritchard JK. Noisy Splicing Drives mRNA Isoform
1356 Diversity in Human Cells. PLoS Genetics. 2010;6: e1001236.
1357 doi:10.1371/journal.pgen.1001236
1358 53. Saudemont B, Popa A, Parmley JL, Rocher V, Blugeon C, Necsulea A, et al. The
1359 fitness cost of mis-splicing is the main determinant of alternative splicing patterns. Genome
1360 Biology. 2017;18: 208. doi:10.1186/s13059-017-1344-6
1361 54. Tress ML, Abascal F, Valencia A. Most Alternative Isoforms Are Not Functionally
1362 Important. Trends in Biochemical Sciences. 2017;42: 408–410.
1363 doi:10.1016/j.tibs.2017.04.002
1364 55. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative
1365 splicing. Nature. 2010;463: 457–463. doi:10.1038/nature08909
1366 56. McGlincy NJ, Smith CWJ. Alternative splicing resulting in nonsense-mediated mRNA
1367 decay: what is the meaning of nonsense? Trends in Biochemical Sciences. 2008;33: 385–393.
1368 doi:10.1016/j.tibs.2008.06.001
1369 57. Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. Unproductive splicing of
1370 SR genes associated with highly conserved and ultraconserved DNA elements. Nature.
1371 2007;446: 926–929. doi:10.1038/nature05676
1372 58. Lareau LF, Brenner SE. Regulation of Splicing Factors by Alternative Splicing and
1373 NMD Is Conserved between Kingdoms Yet Evolutionarily Flexible. Molecular Biology and
1374 Evolution. 2015;32: 1072–1079. doi:10.1093/molbev/msv002

1375 59. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome
1376 evolution. Science. 1980;284: 601–603.
1377 60. Orgel LE, Crick FHC. Selfish DNA: the ultimate parasite. Science. 1980;284: 604–
1378 607.
1379 61. Werren JH. Selfish genetic elements, genetic conflict, and evolutionary innovation.
1380 Proceedings of the National Academy of Sciences. 2011;108: 10863–10870.
1381 doi:10.1073/pnas.1102343108
1382 62. Krenek S, Berendonk TU, Petzoldt T. Thermal performance curves of Paramecium
1383 caudatum: a model selection approach. Eur J Protistol. 2011;47: 124–137.
1384 doi:10.1016/j.ejop.2010.12.001
1385 63. Lhuillier-Akakpo M, Guérin F, Frapporti A, Duharcourt S. DNA deletion as a
1386 mechanism for developmentally programmed centromere loss. Nucleic Acids Res. 2016;44:
1387 1553–1565. doi:10.1093/nar/gkv1110
1388 64. Cummings DJ. Isolation and partial characterization of macro- and micronuclei from
1389 Paramecium aurella. J Cell Biol. 1972;53: 105–115. doi:10.1083/jcb.53.1.105
1390 65. Freiburg M. Isolation and characterization of macronuclei of Paramecium caudatum
1391 infected with the macronucleus-specific bacterium Holospora obtusa. J Cell Sci. 1985;73:
1392 389–398.
1393 66. Galbraith DW, Harkins KR, Maddox JM, Ayres NM, Sharma DP, Firoozabady E.
1394 Rapid Flow Cytometric Analysis of the Cell Cycle in Intact Plant Tissues. Science. 1983;220:
1395 1049–1051. doi:10.1126/science.220.4601.1049
1396 67. Bourge M, Brown SC, Siljak-Yakovlev S. Flow cytometry as tool in plant sciences,
1397 with emphasis on genome size and ploidy level assessment. Genetics & Applications. 2018;2:
1398 1–12. doi:10.31383/ga.vol2iss2pp1-12
1399 68. Dolezel J, Bartos J, Voglmayr H, Greilhuber J. Nuclear DNA content and genome size

1400 of trout and human. Cytometry A. 2003;51: 127–128. doi:10.1002/cyto.a.10013
1401 69. Marie D, Brown SC. A cytometric exercise in plant DNA histograms, with 2C values
1402 for 70 species. Biol Cell. 1993;78: 41–51. doi:10.1016/0248-4900(93)90113-s
1403 70. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled
1404 contigs using SSPACE. Bioinformatics. 2011;27: 578–579.
1405 doi:10.1093/bioinformatics/btq683
1406 71. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically
1407 improved memory-efficient short-read de novo assembler. Gigascience. 2012;1: 18.
1408 doi:10.1186/2047-217X-1-18
1409 72. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using
1410 exact alignments. Genome Biol. 2014;15: R46. doi:10.1186/gb-2014-15-3-r46
1411 73. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods.
1412 2012;9: 357–359. doi:10.1038/nmeth.1923
1413 74. Le Mouël A, Butler A, Caron F, Meyer E. Developmentally regulated chromosome
1414 fragmentation linked to imprecise elimination of repeated sequences in paramecia. Eukaryotic
1415 Cell. 2003;2: 1076–1090. doi:10.1128/ec.2.5.1076-1090.2003
1416 75. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an
1417 information aesthetic for comparative genomics. Genome Res. 2009;19: 1639–1645.
1418 doi:10.1101/gr.092759.109
1419 76. Foissac S, Gouzy J, Rombauts S, Mathe C, Amselem J, Sterck L, et al. Genome
1420 Annotation in Plants and Fungi: EuGene as a Model Platform. Current Bioinformatics.
1421 2008;3: 87–97.
1422 77. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO:
1423 assessing genome assembly and annotation completeness with single-copy orthologs.
1424 Bioinformatics. 2015;31: 3210–3212. doi:10.1093/bioinformatics/btv351

1425 78. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of
1426 occurrences of k-mers. Bioinformatics. 2011;27: 764–770. doi:10.1093/bioinformatics/btr011
1427 79. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.
1428 BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421.
1429 doi:10.1186/1471-2105-10-421
1430 80. Miele V, Penel S, Duret L. Ultra-fast sequence clustering from similarity networks
1431 with SiLiX. BMC Bioinformatics. 2011;12: 116. doi:10.1186/1471-2105-12-116
1432 81. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:
1433 improvements in performance and usability. Mol Biol Evol. 2013;30: 772–780.
1434 doi:10.1093/molbev/mst010
1435 82. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective
1436 stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol.
1437 2015;32: 268–274. doi:10.1093/molbev/msu300
1438 83. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, et al. Global trends of
1439 whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature. 2006;444:
1440 171–178. doi:10.1038/nature05230
1441 84. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and
1442 ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007;56: 564–577.
1443 doi:10.1080/10635150701472164
1444 85. Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, et al. RevBayes:
1445 Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-
1446 Specification Language. Syst Biol. 2016;65: 726–736. doi:10.1093/sysbio/syw021
1447 86. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs.
1448 Bioinformatics. 2013;29: 2487–2489. doi:10.1093/bioinformatics/btt403
1449 87. Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. De novo

1450 assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with
1451 dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito
1452 (Aedes aegypti). Genome Biol Evol. 2015;7: 1192–1205. doi:10.1093/gbe/evv050
1453
1454

1455 ## Supporting information

1456

1457 **S1 Fig. Multi-gate flow cytometry strategy for sorting the MICs.**

1458 GFP, DAPI-positive MICs from *P. sonneborni* vegetative cells transformed with the *P.*
1459 *tetraurelia CENH3a-GFP* transgene [63] were sorted based on size, granularity, DAPI staining
1460 and GFP signal (see Materials and Methods). P4 and P8 were sorted separately. Based on
1461 quality control by flow imaging (Imagestream) indicating 97% purity, the two samples P4 and
1462 P8, which represent 1.91% of total events, were combined for DNA extraction and sequencing.
1463 Two populations are visualized and likely correspond to 2n and 4n MICs.

1464

1465 **S2 Fig. Estimating the proportion of MIC and MAC DNA in the sample based on IES**
1466 **retention score.**

1467 The histograms on the left show the k-mer depth profiles. The peak at the origin can be
1468 attributed to sequencing errors (k-mers that occur only once or a few times). The position of
1469 the largest peak beyond the origin corresponds to k-mers present once in the genome and
1470 provides the sequencing depth. As *Paramecium aurelia* genomes have undergone whole
1471 genome duplications, there are a significant number of k-mers at 2X and even 4X the
1472 sequencing depth arising from genes (or regions of genes) present in 2 or 4 copies, clearly
1473 visible for *P. octaurelia* and *P. primaurelia*. The profile for *P. tetraurelia* however has a first
1474 peak (MIC sequences that occur once) at 31X followed by a larger peak that is not at the 2X
1475 position as it arises because of MAC DNA contamination. The column on the right shows
1476 histograms of IES retention scores. Only the *P. tetraurelia* sample is significantly contaminated
1477 by MAC DNA: the average IES retention score of 0.4 indicates 40% MIC and 60% MAC DNA
1478 in this sample.

1479

1480 **S3 Fig. Comparison of cytometry and k-mer MIC genome size estimates.**

1481 Flow cytometry estimates of DNA content of micronuclei and k-mer counting estimates of
1482 genome size are described in Materials and Methods. In order to show all of the data, both axes
1483 of the graph are log-transformed. Simple linear regression was carried out on the untransformed
1484 data with R. The linear model that fits the data is presented as a dashed blue line; $R^2 = 0.99$, p-
1485 value = 1.3×10^{-09}.

1486

1487 **S4 Fig. Repeat content of *P. caudatum* MIC genome.**

1488 (A) Abundance of repeat families identified by DNAPipeTE in *P. caudatum* strain My43c3d.
1489 The repeat content of the *P. caudatum* MIC genome was analyzed with DNAPipeTE [87], using
1490 a sample of 3,500,000 sequence reads (corresponding to a read depth of ~0.5X). DNAPipeTE
1491 identified 67 repeat families that collectively constitute 83% of the MIC genome. Among them,
1492 there are two major satellite repeats Sat1 and Sat2, which represent respectively 42% and 29%
1493 of the MIC genome.

1494 (B) Sequences of the two major satellite repeats Sat1 and Sat2 in *P. caudataum* My43c3d (332
1495 bp and 449 bp long). These two satellite repeats share homology over a ~200 bp-long region.
1496 Primer sequences used for specific PCR amplification of each repeat are indicated in bold.

1497 (C) Detection of Sat1 and Sat2 *in P. caudatum* strains. Whole cell genomic DNA was used to
1498 perform duplex PCR with a set of primers located within each repeat (Sat1 or Sat2, in bold
1499 panel B) and another set of primers within the 18S ribosomal DNA as a loading control. The
1500 expected size of the 18SrDNA PCR product was 301 bp using primers 18S_F953:
1501 AGACGATCAGATACCGTCGTAG and 18S_R1300: CACCAACTAAGAACGGCCATGC.
1502 L: 1-kb NEB ladder. Neg.: negative control (no DNA). Sat1 was amplified with primers
1503 comp2975_F1: TTGTGCTGTAGGGCTCAATAAT and comp2975_R1:
1504 CTCAAAATTCGACGCTGACAA at the expected size (198 bp) in *the P. caudatum* clade B

1505 strains tested (My43c3d; C033; C083; C131; C147). The repeat could not be amplified in *P.*
1506 *caudatum* DNA from clade A strains (C023; C065; C104; C119), from strain C026 or from
1507 strain Indo_1.6I.
1508 Sat2 was amplified with primers comp5240_F1: TGCTGCTGATTTTGGATCTCG and
1509 comp5240_R1: CCGAGAACGGCCATTACAAG at the expected size (168 bp) in the *P.*
1510 *caudatum* clade B strains tested (My43c3d; C033; C083; C131; C147). The repeat could not
1511 be amplified in *P. caudatum* DNA from clade A strains (C023; C065; C104; C119), from strain
1512 C026 or from strain Indo_1.6I.
1513
1514 **S5 Fig. Impact of sequencing depth on IES detection**. To assess the impact of sequencing
1515 depth on the sensitivity of IES detection, we subsampled sequence reads from the *P. tetraurelia*
1516 MIC dataset so as to obtain subsets of lower depth (from 2x to 30x), on which we applied the
1517 same IES detection procedure. (A) Number of detected IESs vs. sequencing depth. (B-H)
1518 Length distributions of IESs detected within each subset (the percentage of IESs in each peak
1519 is indicated). Sequencing depth affects the number of detected IESs (for depths <15x), but not
1520 their length distribution.
1521
1522 **S6 Fig. Example of a MAC-variable region**
1523 Circular representation of one scaffold of ~730 kb. The tracks from the exterior to the interior
1524 of the circle: G+C content of 100nt sliding windows (black), MAC DNA-seq depth (purple),
1525 the density in predicted non-coding genes (orange), RNA-seq depth (red) and the density of
1526 detected telomerisation sites (green). The external blue arc shows the region identified as being
1527 MAC-variable. These regions were determined by an automatic pipeline (see Materials and
1528 Methods), then adjusted by eye for each scaffold.
1529

1530 **S7 Fig. Intragenic IES density vs. gene expression level**.

1531 Expression levels (RPKM) were measured with RNAseq datasets from vegetative cells. For
1532 each species, expressed genes were classified into 10 bins of equal sample size according to
1533 their expression level, and IES density computed within each bin. Non-expressed genes (6.6%
1534 of the entire dataset) were excluded. (A) *Paramecium aurelia* species. (B) *P. caudatum*.

1535

1536 **S8 Fig. Measuring the rate of IES gain or loss along the species phylogeny**.

1537 To illustrate our methodology, we show here an example of a gene family with 3 genes, two
1538 from *P. sonneborni* (pson1, pson2) and one from *P. sexaurelia* (psex1). Two IES loci are found
1539 in this family (A, B). The probability of presence of an IES (estimated by Bayesian ancestral
1540 state reconstruction - see Materials and Methods) is indicated by shaded circles for each locus
1541 at each node of the gene phylogeny. We focus here on the branch of the species tree leading
1542 from the common ancestor of *P. sexaurelia* and *P. sonneborni* to the leaf node of *P. sonneborni*
1543 (the red branch in the species tree, shown in the insert). The length of this branch (*b*), is taken
1544 as a proxy for time. Because of a duplication event, this branch of the species tree corresponds
1545 to two paths in the gene tree (k=2). To estimate the IES gain rate, we calculate for each path the
1546 sum of increase in the probability of presence of an IES, for all IES loci (p^+). Along the first
1547 path (from the root to pson1), we have p^+A1=0.5 and p^+B1=0. Along the second path (from the
1548 root to pson2), we have p^+A2=0.5 and p^+B2=0. The average gain rate along all paths, per unit
1549 of time and per bp, is thus given by $G=(p^+A1 + p^+B1 + p^+A2 + p^+B2)/(k \times b \times n_g)$, where n_g is
1550 the number of well aligned sites in the gene family alignment (i.e. the number of sites where
1551 the presence of co-orthologous IESs can be assessed). Similarly, to estimate the IES loss rate,
1552 we calculate for each path the sum of decrease in the probability of presence of an IES, for all
1553 IES loci (p^-). Along the first path (from the root to pson1), we have p^-A1=0 and p^-B1=0.4. Along
1554 the second path (from the root to pson2), we have p^-A2=0 and p^-B2=0.4. The average gain rate

1555 along all paths, per unit of time and per bp, is thus given by $L=(p^{-}A1 + p^{-}B1 + p^{-}A2 + p^{-}B2)/(k$
1556 $\times b \times I)$, where I is the number of IES loci in the gene family (here I=2).
1557
1558 **S9 Fig. Dating events of IES insertion/loss**.
1559 (A) To date events of IES loss or gain, it is first necessary to identify IESs that derive from a
1560 single ancestral insertion event (co-orthologous IESs). For this, we aligned coding sequences
1561 (based on the protein alignment) and mapped the position of IESs: IESs located at the exact
1562 same position within a codon were assumed to be co-orthologous. We then used the reconciled
1563 gene tree to map events in the species phylogeny, using a maximum likelihood approach (see
1564 Materials and Methods). The example shown here corresponds to a gene family encoding a
1565 putative RNA 3'-terminal phosphate cyclase (PTET.51.1.P0920097, POCTA.138.1.P0960088,
1566 PBIA.V1_4.1.P01950012, PTRED.209.2.P71800001293600070, PPENT.87.1.P1090087,
1567 PPRIM.AZ9-3.1.P0020612, PSON.ATCC_30995.1.P0860097, PSEX.AZ8_4.1.P0910047,
1568 PCAU.43c3d.1.P00760109). The positions of IESs are indicated by red rectangles. (B) The
1569 presence of IESs (red bars) within each of these genes is indicated with regard to the species
1570 phylogeny. Six distinct IESs were identified in this gene family: IES2 is shared by all species
1571 and therefore predates the divergence between *P. caudatum* and the *aurelia* clade; IES4 most
1572 probably corresponds to a gain in the *P. sexaurelia* lineage; IES5 and IES6 predate the
1573 divergence of the *aurelia* clade and have been subsequently lost in the *P. tetraurelia/P.*
1574 *octaurelia* lineage; IES1 might correspond to a gain at the base of the *aurelia* clade or a loss in
1575 the *P. caudatum* lineage (and vice versa for IES3).
1576
1577 **S10 Fig. Prevalence of weak IESs**.
1578 (A) Proportion of weak IESs (i.e. IESs with a retention frequency ≥10% in WT vegetative cells)
1579 among IESs located in different genomic compartments. The proportion of weak IESs among

1580 intergenic IESs was compared to that of IESs located in introns or coding regions, by a Chi-
1581 squared test. (B) Proportion of weak IESs according to the age of IESs (for IESs located in
1582 coding regions): New = species-specific IES; Old = IES predating the divergence between *P.*
1583 *caudatum* and the *aurelia* lineage. The number of new IESs is indicated for each species. The
1584 proportion of weak IESs among new IESs was compared to that of older ones, by a Chi-squared
1585 test. Species codes: pso: *P. sonneborni*, ptr: *P. tredecaurelia*, pte: *P. tetraurelia*, pbi: *P.*
1586 *biaurelia*, poc: *P. octaurelia*, pse: *P. sexaurelia*, ppr: *P. primaurelia*, ppe: *P. pentaurelia*, pca:
1587 *P. caudatum*. (*: p-value < 0.05; **: p-value < 1e-3; ***: p-value < 1e-6).
1588
1589 **S11 Fig. Length distribution of IESs according to their age**.
1590 Comparison of the length distribution of IESs according to their age (for the subset of datable
1591 IESs located in coding regions). The age of an IES site is defined as in Fig 3. Results for other
1592 species are shown in Fig 4.
1593
1594 **S12 Fig. Genomic distribution of IESs according to their length**.
1595 Green bars indicate the percentage of IESs located in each compartment of the MAC genome
1596 (introns, protein-coding regions, intergenic regions) for each species. Grey bars indicate the
1597 percentage of the MAC genome in each compartment. (A) Long IESs (>100 bp). (B) Short IESs
1598 (<35 bp). For each species, the relative proportion of IESs in the three compartments (intron,
1599 intergenic, coding regions) was compared for short IESs (<35 bp) vs long IESs (> 100 bp) by
1600 a Chi-squared test (p-value $<10^{-16}$ in all species).
1601
1602 **S13 Fig. Example of floating IES**.
1603 Comparison of MIC and MAC sequences indicates the presence of an IES at this locus.
1604 However, because of the presence of a repeated motif at the boundaries of the IES (blue arrows),

1605 it is not possible to determine which of the two possible segments (IES-1 in black or IES-2 in
1606 red) is actually excised *in vivo*. Such IESs that cannot be unambiguously positioned are called
1607 'floating IESs'. They represent 6.8% of the 400,254 IESs detected across all species. In the vast
1608 majority of cases (86%) the alternative locations of floating IESs differ by only two bp (as in
1609 the example shown here), and there are less than 1% of floating IESs for which the uncertainty
1610 in IES position exceeds 5 bp.
1611
1612 **S1 Table. MIC genomes: sequencing data and size estimates.** (a) Highly enriched sorted
1613 MIC (97-99%) are contaminated with the highly polyploid MAC. For instance, 97% of MIC
1614 purity led to 60% of MAC DNA contamination in the case of P. tetraurelia (see S2 Fig). MIC
1615 genome size (in Mb) was estimated based on (b) flow cytometry analysis and (c) k-mer counts.
1616 Size estimation before correction, based on MAC contamination, is indicated in parentheses
1617 (see Material and Methods).
1618
1619 **S2 Table. MAC genome assemblies used in this study.**
1620 The assemblies of the four MAC genomes sequenced in the course of this project include both
1621 'constitutive MAC' regions (i.e. regions that are always retained in the MAC) and 'MAC-
1622 variable regions' (i.e. regions that are mostly restricted to the MIC, but that are retained at low
1623 frequency in MAC nuclei). The size and content of these two types of regions are indicated.
1624 The completeness of MAC genome assemblies was assessed using BUSCO [77].
1625
1626 **S3 Table. Distribution of IESs in different genomic compartments.**
1627 Values in parentheses indicate the proportions expected under the hypothesis of uniform IES
1628 distribution along MAC-destined regions.
1629

1630 **S4 Table. Genomic distribution of copies of mobile IESs.**

1631 Copies of mobile IESs were searched within IES sequences (see Materials and Methods).
1632 Detected copies are divided in two categories: nested copies (i.e. copies inserted within an IES,
1633 but not including the extremities of the IES) and *bona fide* mobile IESs (i.e. copies whose
1634 extremities correspond to the extremities of the IES). This table lists all families for which at
1635 least one species contains ≥ 10 copies of *bona fide* mobile IESs in its genome. The genomic
1636 distribution (in coding regions, introns and intergenic regions) of IESs containing these copies
1637 (nested copies or *bona fide* mobile IESs) is indicated.

1638

1639 **S5 Table. Highly conserved IESs that are transcribed during MAC development and/or**
1640 **associated to genes that are upregulated during MAC development**. The transcription level
1641 of IESs is indicated for different stages during MAC development (S, T0 to T45) and in
1642 vegetative cells (V) [37]. Only *P. tetraurelia* IESs are shown in the table, because this is the
1643 only species for which developmental transcriptome data is available [37]. (*) The IES
1644 pte.MICA.16.324097 (FAM_4968) contains a complete protein-coding gene, which is
1645 expressed during development (HTH CenpB-type DNA-binding domain see Fig 5A). The gene
1646 in which this IES is inserted (PTET.51.1.P0160202) is not specifically expressed during
1647 development.

1648

1649 **S6 Table. Genetic distances between species of the *aurelia* complex.** This matrix reports for
1650 each pair of species, the number of orthologous genes analyzed (above the diagonal), and their
1651 median synonymous divergence (below the diagonal).

1652

1653 **S7 Table. Sequencing data generated for this study.**

1654

1655 **S1 Data. Flow cytometry-based estimations of MIC genome size in *Paramecium*.**

1656

1657 **S2 Data. Locations of MAC-variable regions and MAC assembly curation**

1658 This file provides the positions of MAC-variable regions identified in the MAC assemblies of
1659 *P. octaurelia, P. pentaurelia, P. primaurelia,* and *P. sonneborni*. In addition, it indicates the
1660 positions of putative assembly chimeras that have been identified in *P. octaurelia*, *P.*
1661 *primaurelia* and *P. sexaurelia*.

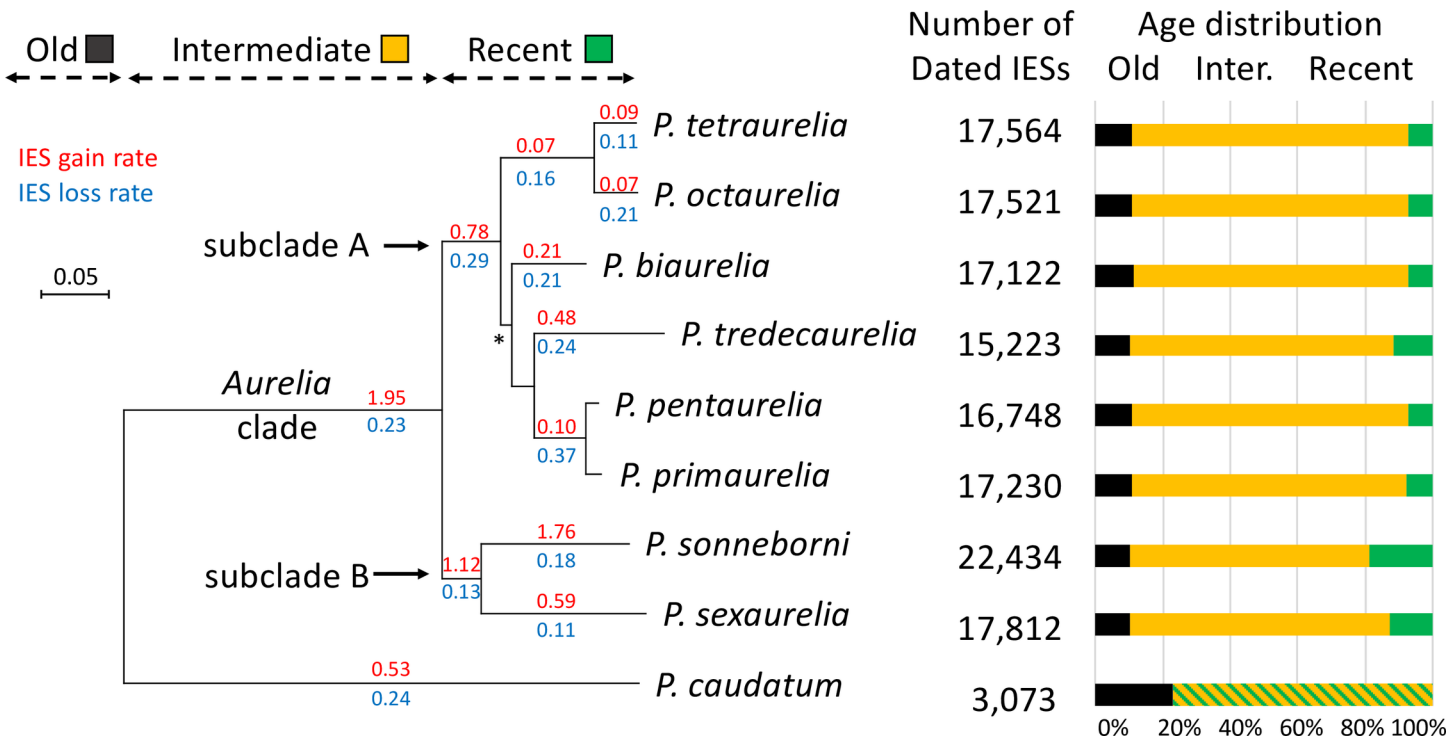

Old
Intermediate
Recent
IES gain rate
IES loss rate
0.05
subclade A
subclade B
Aurelia clade
1.95
0.23
0.78
0.29
0.07
0.16
0.09
0.11
0.07
0.21
0.21
0.21
0.48
0.24
0.10
0.37
1.12
0.13
1.76
0.18
0.59
0.11
0.53
0.24
P. tetraurelia
P. octaurelia
P. biaurelia
P. tredecaurelia
P. pentaurelia
P. primaurelia
P. sonneborni
P. sexaurelia
P. caudatum
Number of Dated IESs
Age distribution
Old
Inter.
Recent
17,564
17,521
17,122
15,223
16,748
17,230
22,434
17,812
3,073
0%
20%
40%
60%
80%
100%

Figure 1

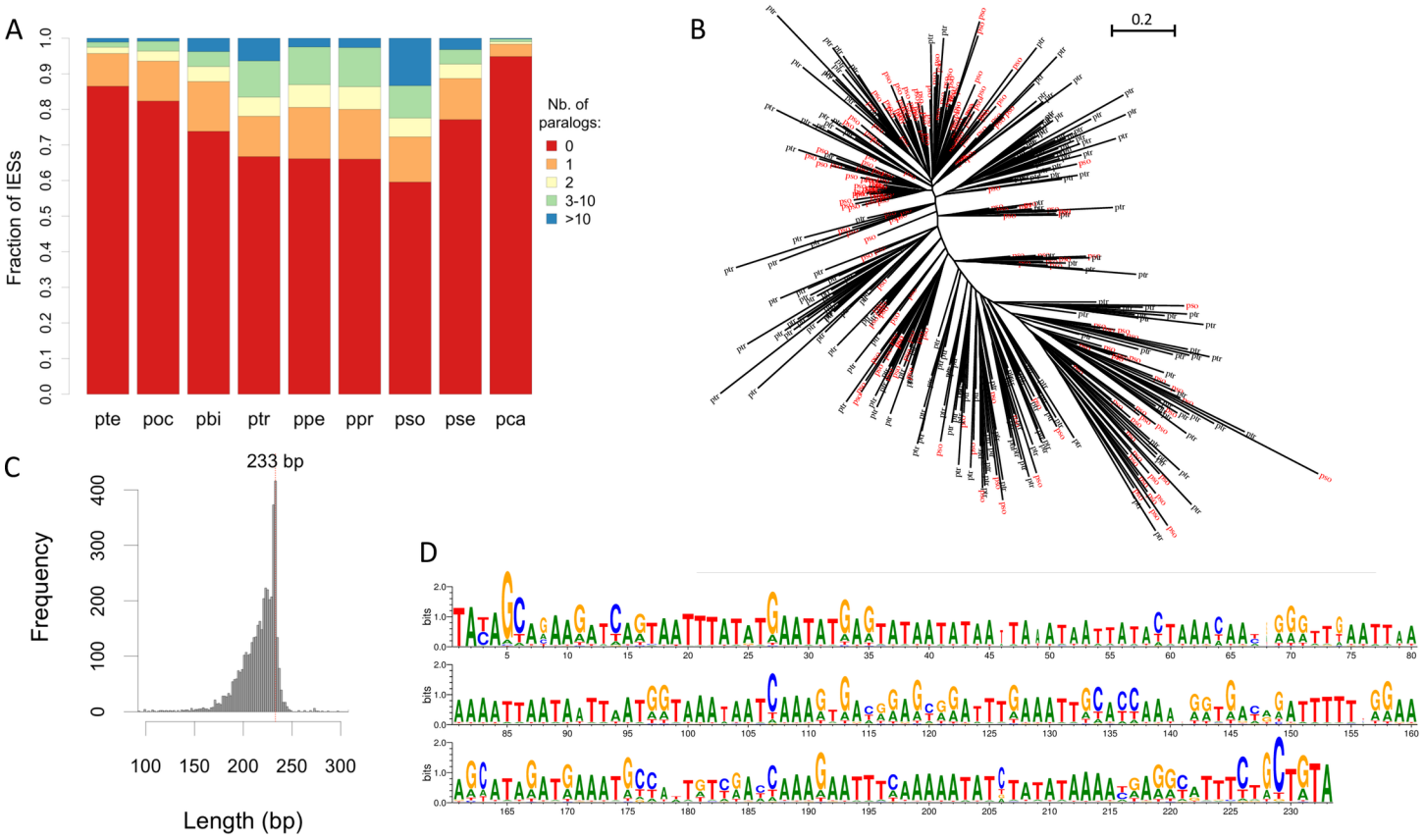

A
Fraction of IESs
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
pte
poc
pbi
ptr
ppe
ppr
pso
pse
pca
Nb. of paralogs:
0
1
2
3-10
>10
B
0.2
C
233 bp
Frequency
0
100
200
300
400
100
150
200
250
300
Length (bp)
D
bits

Figure 2

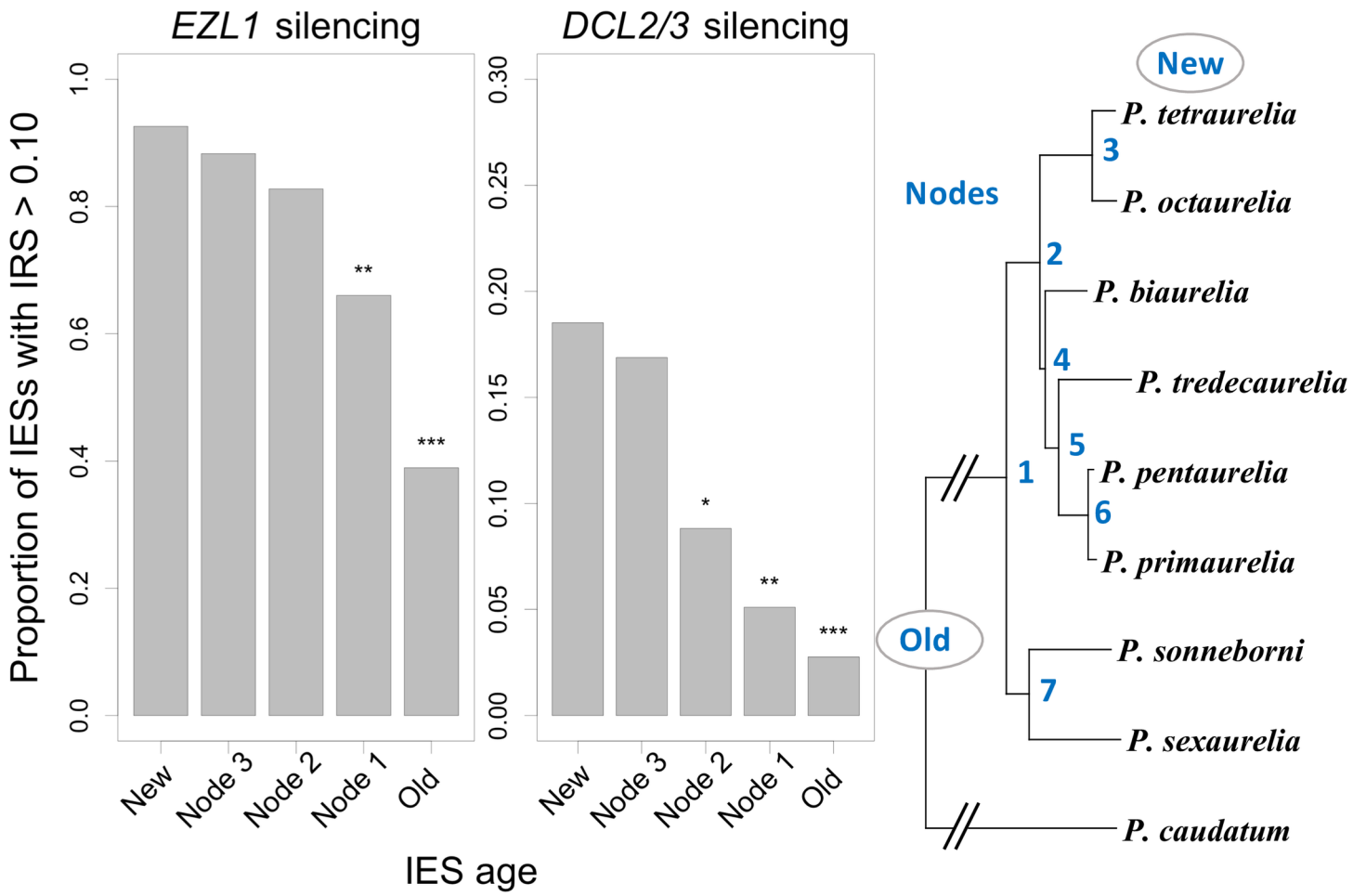

EZL1 silencing
DCL2/3 silencing
Proportion of IESs with IRS > 0.10
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
**
***
*
**
***
New
Node 3
Node 2
Node 1
Old
IES age
Nodes
New
Old
1
2
3
4
5
6
7
P. tetraurelia
P. octaurelia
P. biaurelia
P. tredecaurelia
P. pentaurelia
P. primaurelia
P. sonneborni
P. sexaurelia
P. caudatum

Figure 3

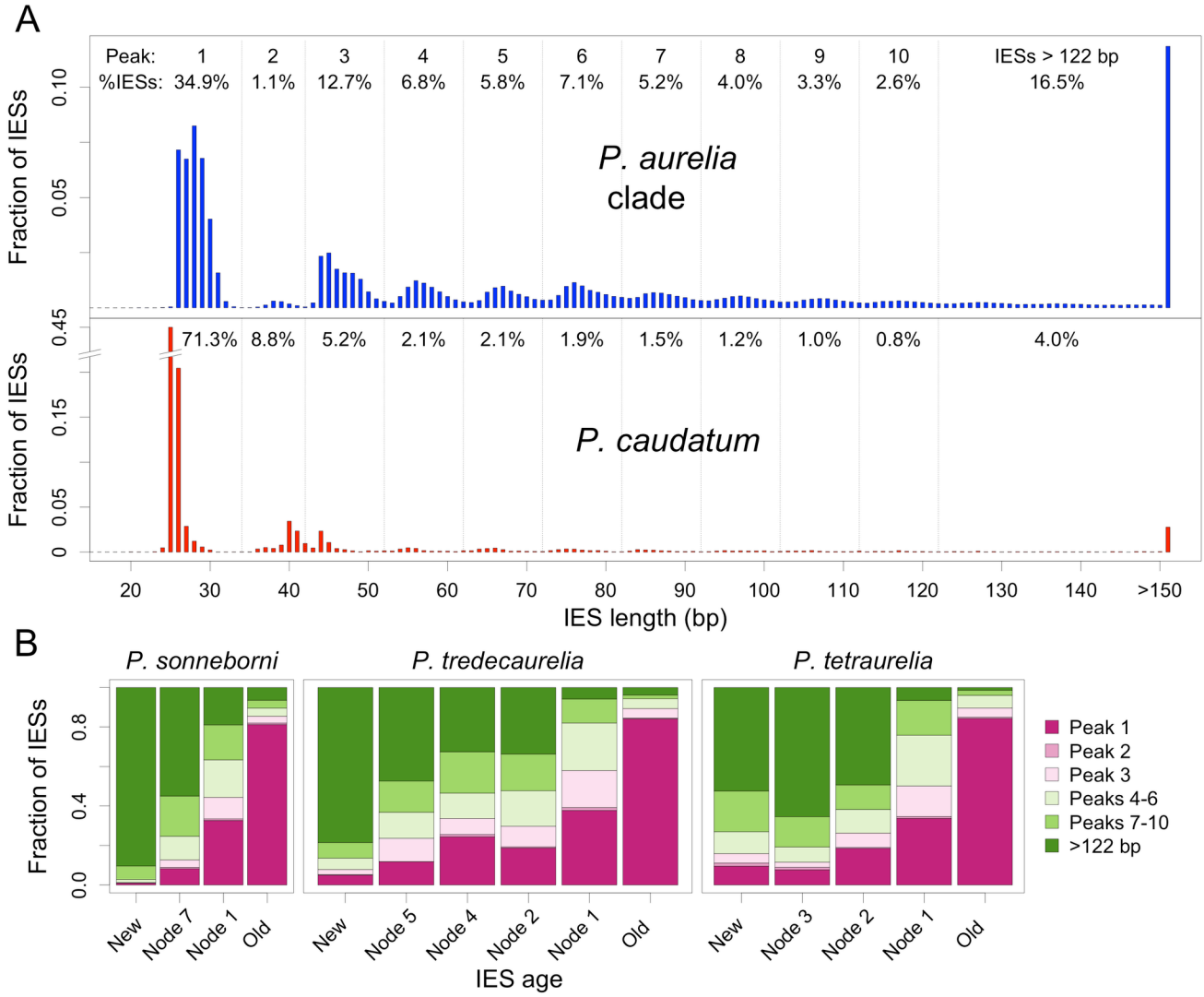

A
Peak: 1 2 3 4 5 6 7 8 9 10 IESs > 122 bp
%IESs: 34.9% 1.1% 12.7% 6.8% 5.8% 7.1% 5.2% 4.0% 3.3% 2.6% 16.5%
P. aurelia clade
71.3% 8.8% 5.2% 2.1% 2.1% 1.9% 1.5% 1.2% 1.0% 0.8% 4.0%
P. caudatum
Fraction of IESs
IES length (bp)
B
P. sonneborni
P. tredecaurelia
P. tetraurelia
Fraction of IESs
IES age
New
Node 7
Node 1
Old
Node 5
Node 4
Node 2
Node 3
Peak 1
Peak 2
Peak 3
Peaks 4-6
Peaks 7-10
>122 bp

Figure 4

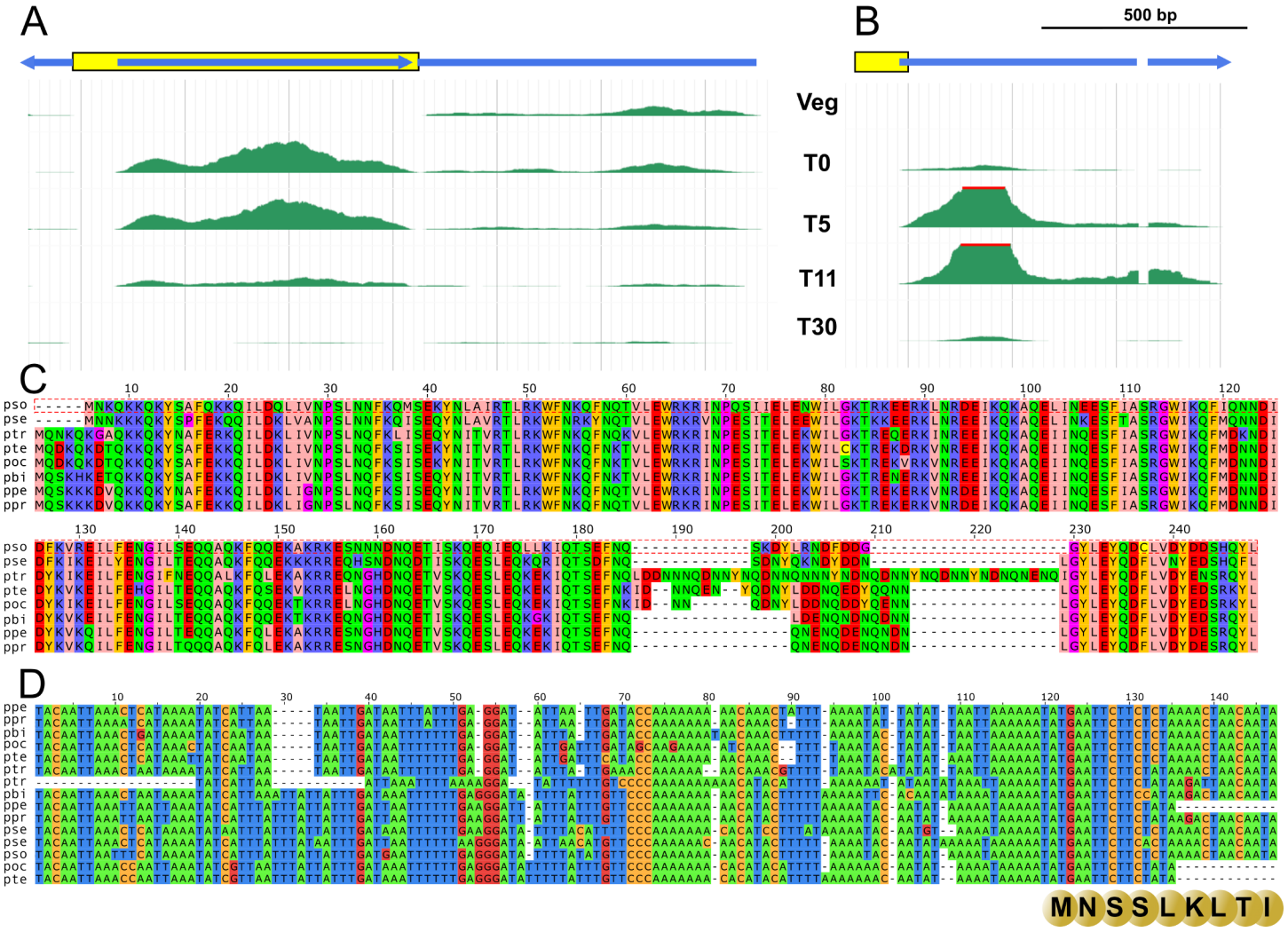
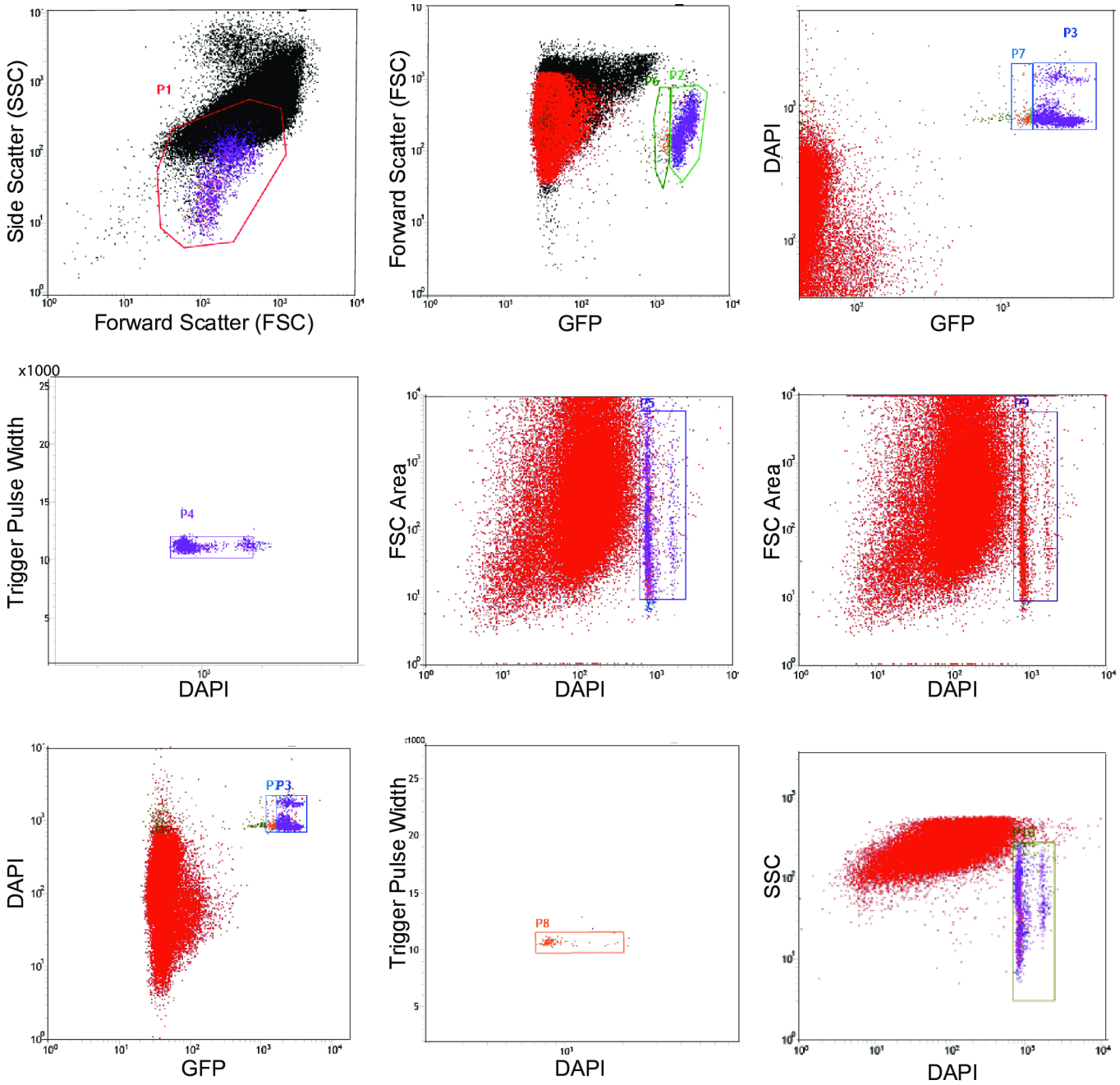


Figure 5



Populations	Events	% Total	% Parent
All Events	100,000	100.00%	####
P1	75,978	75.98%	75.98%
P10	2,465	2.46%	3.24%
P2	2,177	2.18%	88.32%
P3	2,119	2.12%	97.34%
P5	1,868	1.87%	88.15%
P4	1,823	1.82%	97.59%
P6	131	0.13%	5.31%
P7	113	0.11%	86.26%
P9	96	0.10%	84.96%
P8	92	0.09%	95.83%

Figure S1

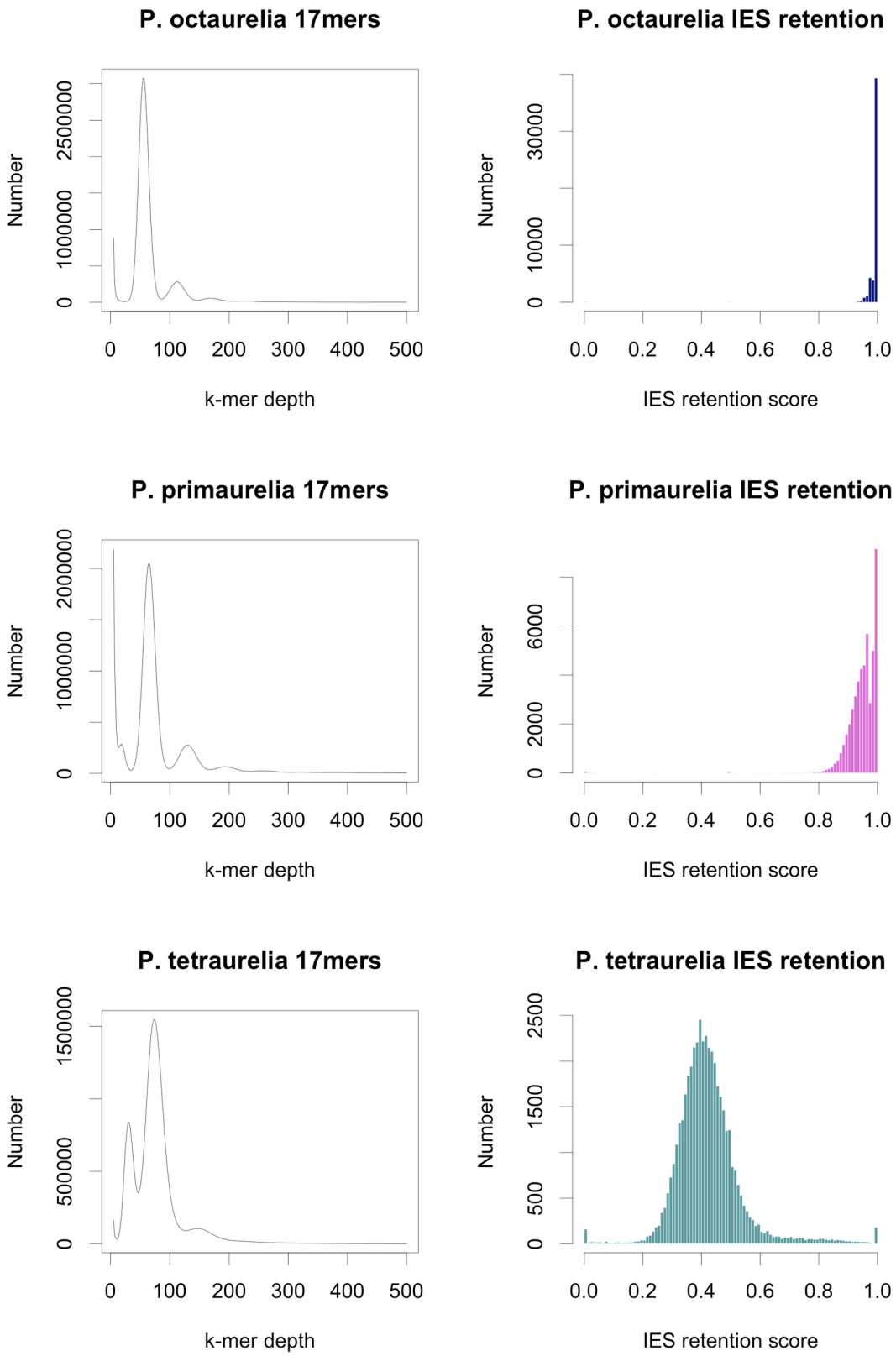

P. octaurelia 17mers
Number
0
1000000
2500000
0
100
200
300
400
500
k-mer depth
P. octaurelia IES retention
Number
0
10000
30000
0.0
0.2
0.4
0.6
0.8
1.0
IES retention score
P. primaurelia 17mers
Number
0
1000000
2000000
0
100
200
300
400
500
k-mer depth
P. primaurelia IES retention
Number
0
2000
6000
0.0
0.2
0.4
0.6
0.8
1.0
IES retention score
P. tetraurelia 17mers
Number
0
500000
1500000
0
100
200
300
400
500
k-mer depth
P. tetraurelia IES retention
Number
0
500
1500
2500
0.0
0.2
0.4
0.6
0.8
1.0
IES retention score

Figure S2

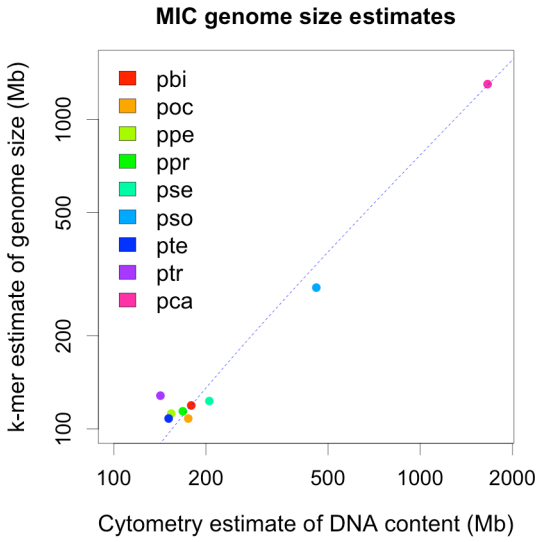
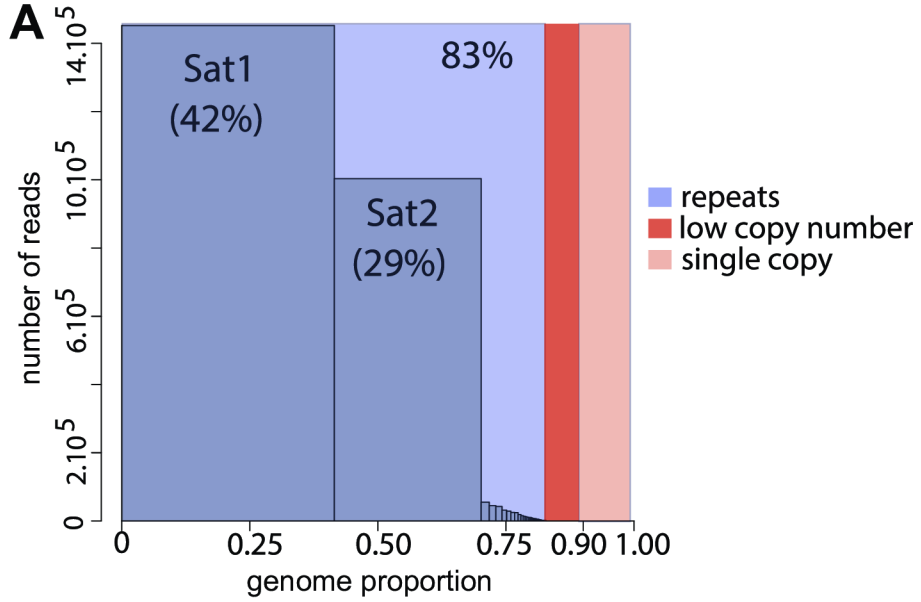

MIC genome size estimates
k-mer estimate of genome size (Mb)
100
200
500
1000
pbi
poc
ppe
ppr
pse
pso
pte
ptr
pca
100
200
500
1000
2000
Cytometry estimate of DNA content (Mb)

Figure S3



B

```
>Sat1: Caudatum_satellite_1 (332 bp)
CATCAAACTTTCTTTTCCTTATGACAAGTCAATTGACATTTTTCGTCAATTTTATGAGGA
GTCATAGGTACCTAATATTTGTGCTGTAGGGCTCAATAATAGCTCATTTTAATGCTGGTA
CAGAGGGCTTTAAAATAGCCATCTCAGTTTTTCTTAATGGAATCTAGATCGCCAGAACGG
GCATTTCATGACTTTTTTAATTATATCATATCGATGGTCATTTCATCTCCAAAGTGTCTA
AAAACACGTTATTCGTTGTCAGCGTCGAATTTTGAGTAAAACTGTCATCTCACTTTTCCT
TAATGGAATTTATTTGAGGAGATCCAATTTCC

>Sat2: Caudatum_satellite_2 (449 bp)
TTTTTGAGATAAAGAACTCGGGTTTTTTTCTATATAATACTGGTACCAAACGCTACAAAT
CGTATGCCTCAGATTTTCTTATTGTTAGCTAAATCCCTGAGAACGCTCATTTTCATCTGG
AAAGTCATCGGGTTTGGTATTTCGCCTGAAAAAATATAAAAAATTTTTGAGATAAAGAAC
TCGGATCTTTTTCTATATAATACTGGTACTAAAAGCTGCAAATCGTATAATCTCGATTTT
CTTAATGGATGCTAAATCCCGAGAACGGCCATTACAAGGGAATTTTAATTTTGCAAAATC
GATGGTTTTTTGATGTCCTACGTTGCTAAAAACATGTTTTATGATATCAGGAGCCCATTT
TACCCCTAATCCACTCTTTTATAGACCTTAATGGAATTTAGTACCCGAGATCCAAAATCA
GCAGCAAATAGCATTTTTCATAGGGAATT
```

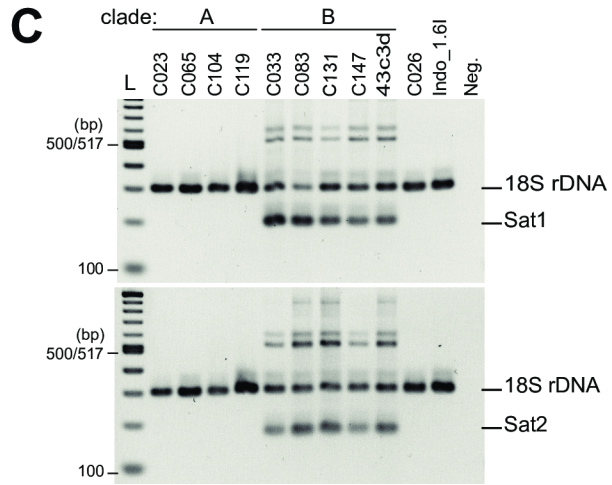


Figure S4

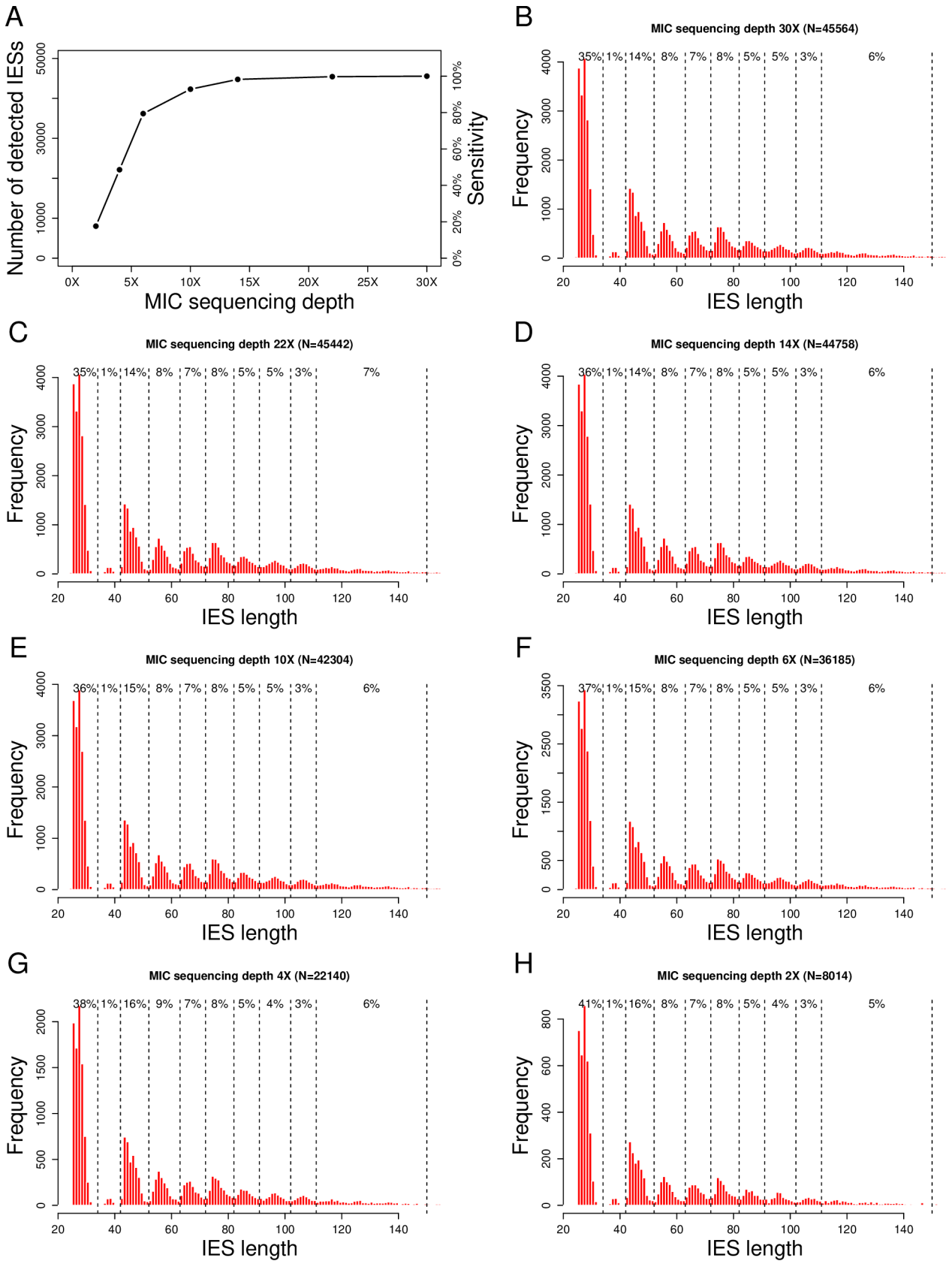

A
Number of detected IESs
Sensitivity
MIC sequencing depth
0X
5X
10X
15X
20X
25X
30X
B
MIC sequencing depth 30X (N=45564)
35% 1% 14% 8% 7% 8% 5% 5% 3% 6%
Frequency
IES length
C
MIC sequencing depth 22X (N=45442)
35% 1% 14% 8% 7% 8% 5% 5% 3% 7%
D
MIC sequencing depth 14X (N=44758)
36% 1% 14% 8% 7% 8% 5% 5% 3% 6%
E
MIC sequencing depth 10X (N=42304)
36% 1% 15% 8% 7% 8% 5% 5% 3% 6%
F
MIC sequencing depth 6X (N=36185)
37% 1% 15% 8% 7% 8% 5% 5% 3% 6%
G
MIC sequencing depth 4X (N=22140)
38% 1% 16% 9% 7% 8% 5% 4% 3% 6%
H
MIC sequencing depth 2X (N=8014)
41% 1% 16% 8% 7% 8% 5% 4% 3% 5%

Figure S5

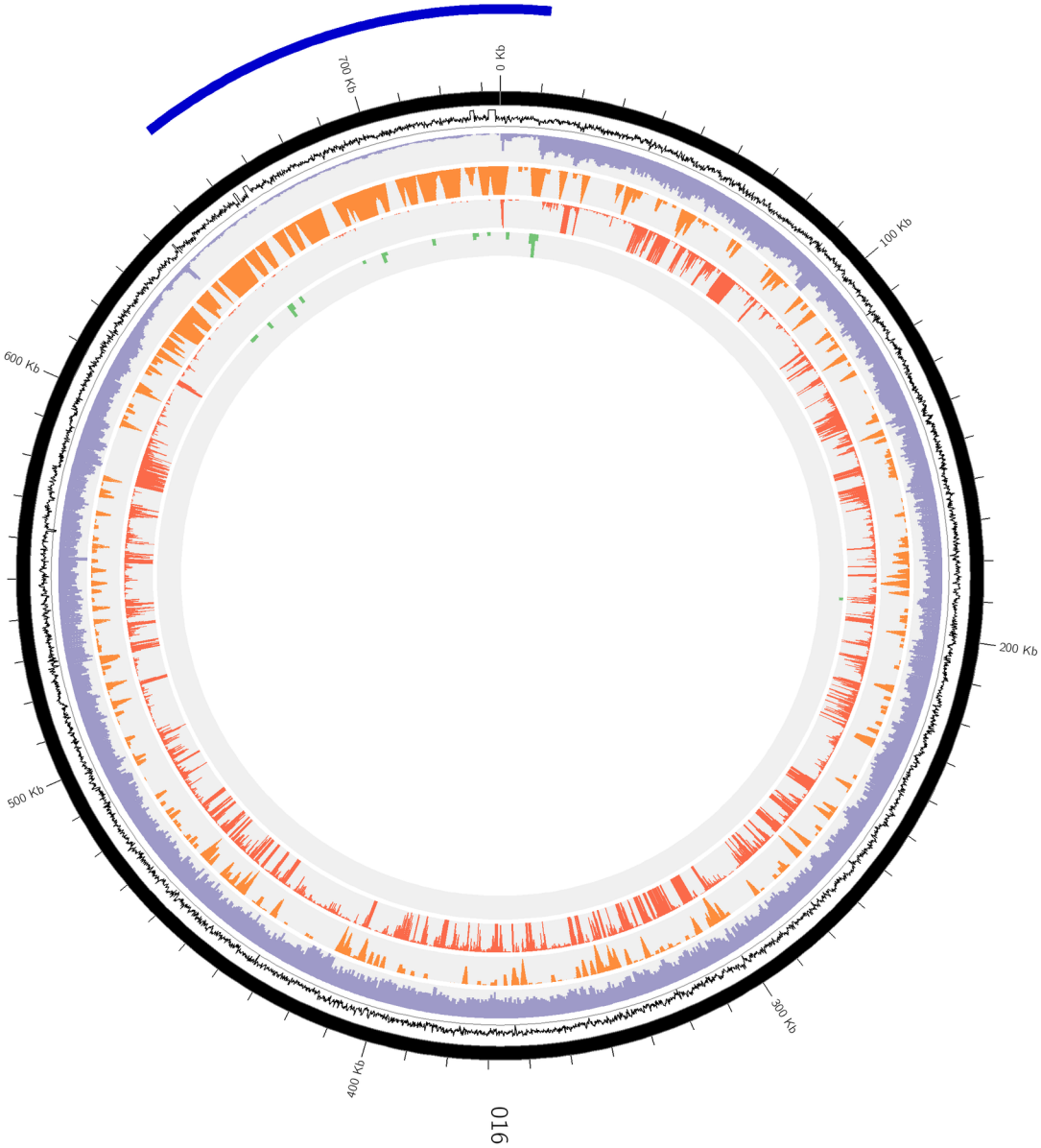


Figure S6

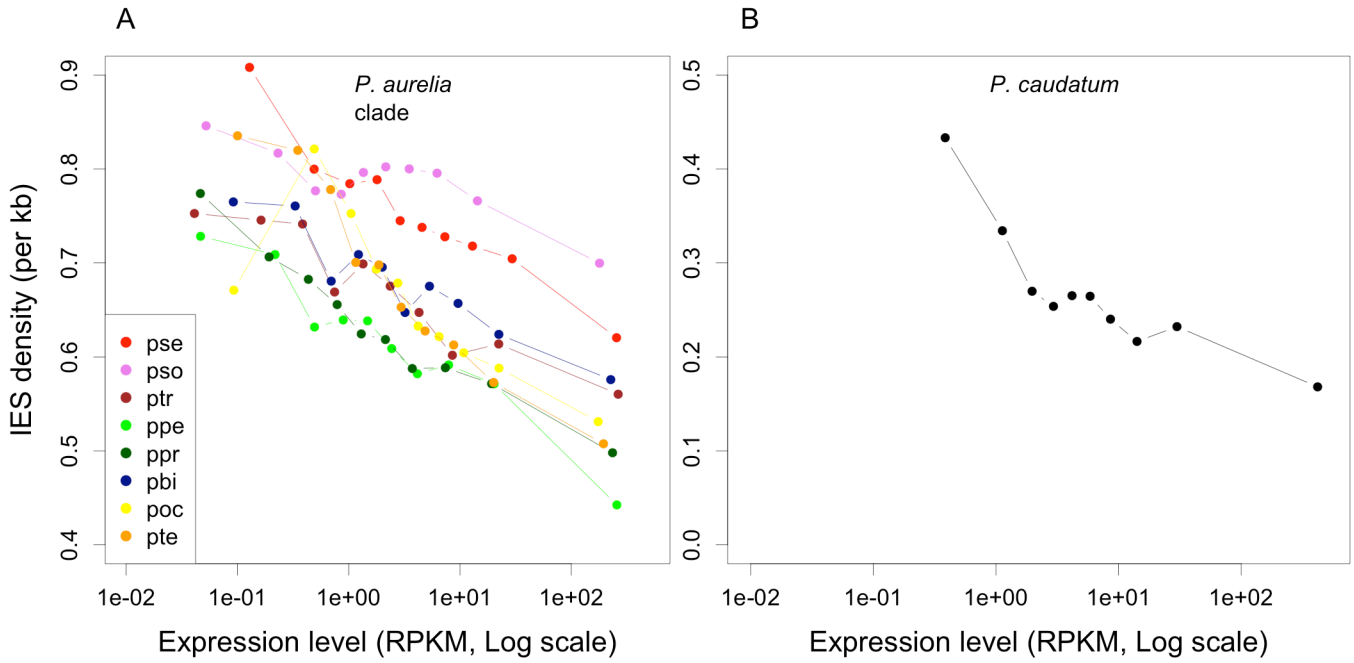

A
B
P. aurelia
clade
P. caudatum
IES density (per kb)
0.4
0.5
0.6
0.7
0.8
0.9
0.0
0.1
0.2
0.3
0.4
0.5
pse
pso
ptr
ppe
ppr
pbi
poc
pte
1e-02
1e-01
1e+00
1e+01
1e+02
Expression level (RPKM, Log scale)

Figure S7

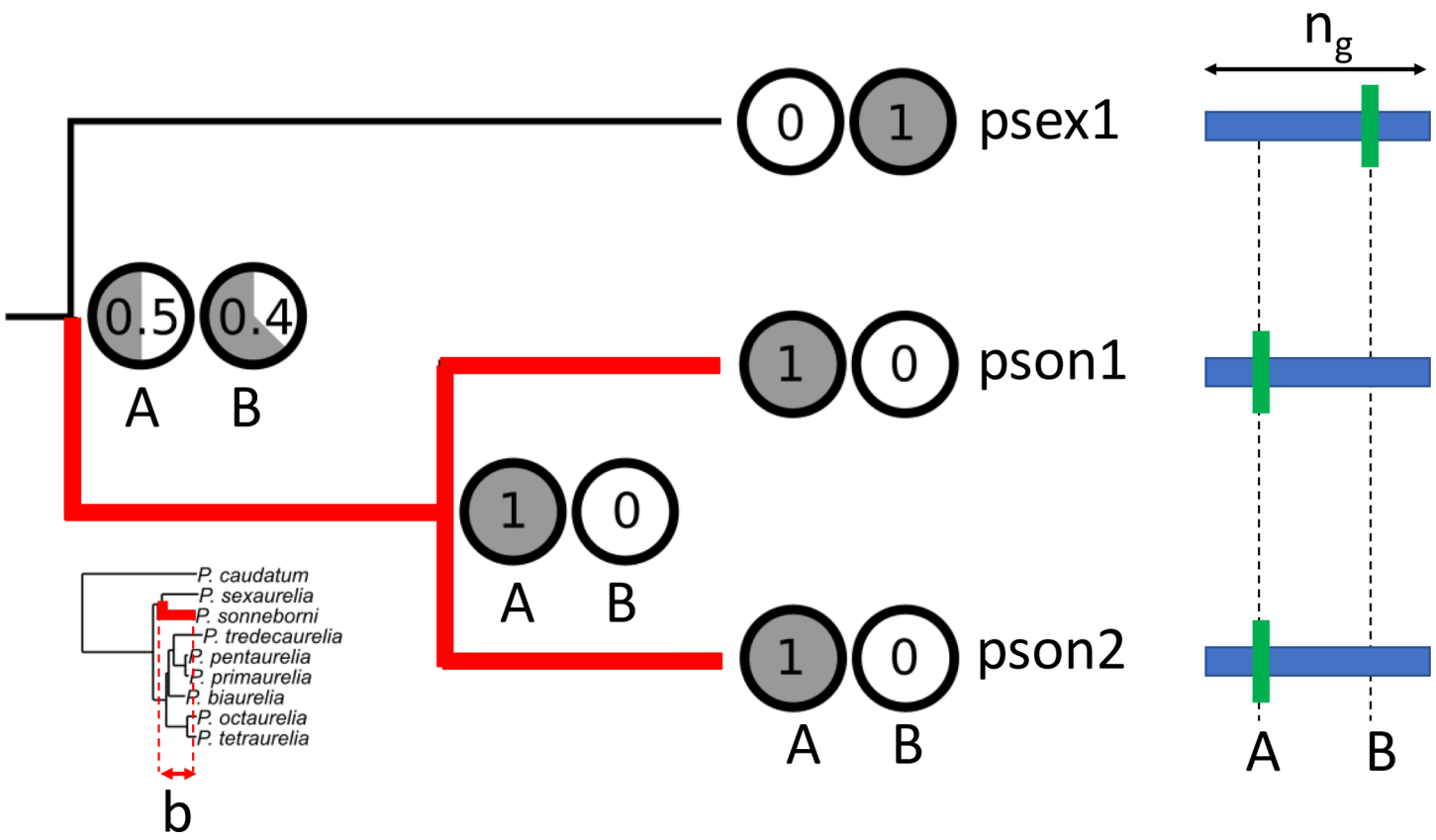

n_g
0
1
psex1
0.5
0.4
A
B
1
0
pson1
1
0
A
B
1
0
pson2
A
B
A
B
P. caudatum
P. sexaurelia
P. sonneborni
P. tredecaurelia
P. pentaurelia
P. primaurelia
P. biaurelia
P. octaurelia
P. tetraurelia
b

Figure S8

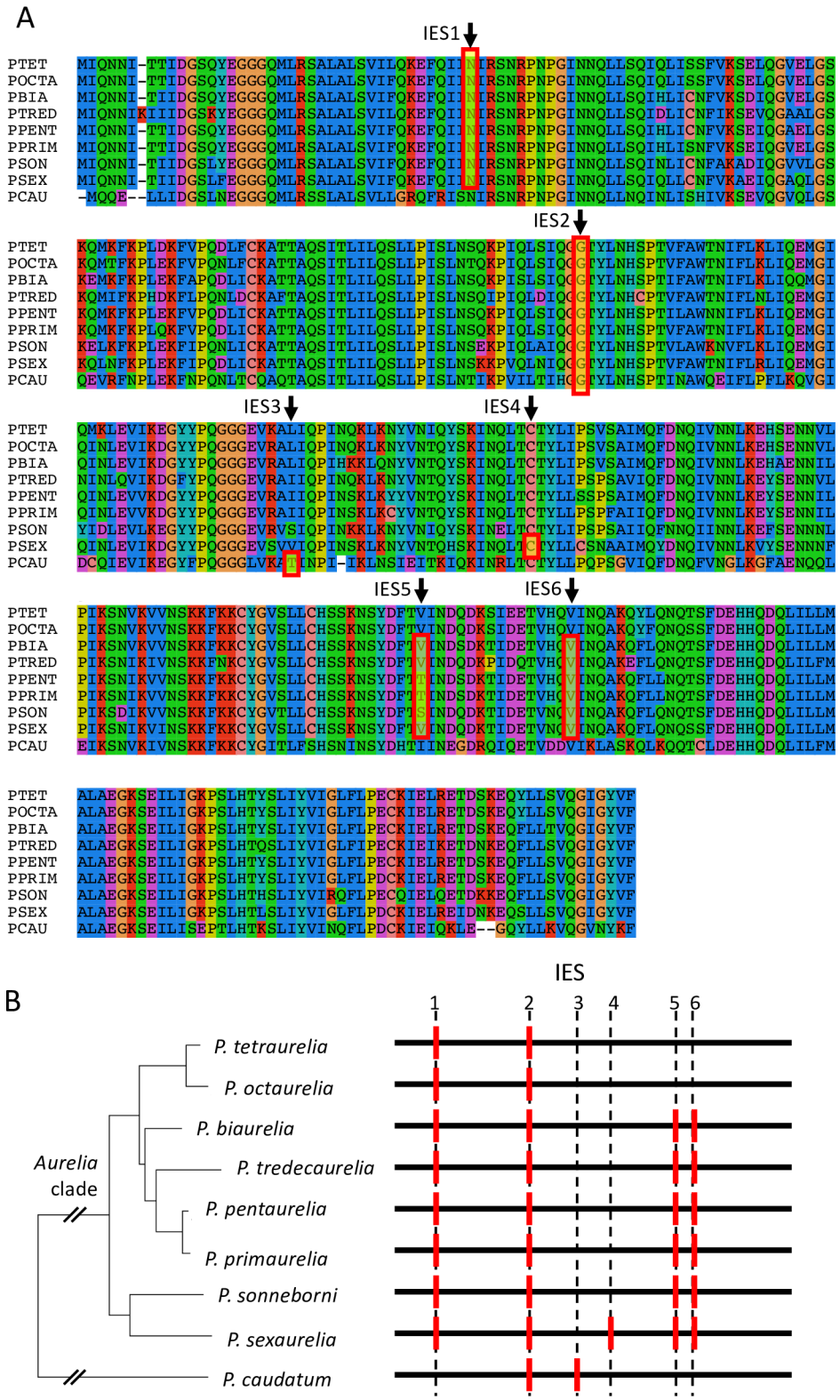

A
IES1
IES2
IES3
IES4
IES5
IES6
PTET
POCTA
PBIA
PTRED
PPENT
PPRIM
PSON
PSEX
PCAU
B
IES
1
2
3
4
5
6
P. tetraurelia
P. octaurelia
P. biaurelia
P. tredecaurelia
P. pentaurelia
P. primaurelia
P. sonneborni
P. sexaurelia
P. caudatum
Aurelia
clade

Figure S9

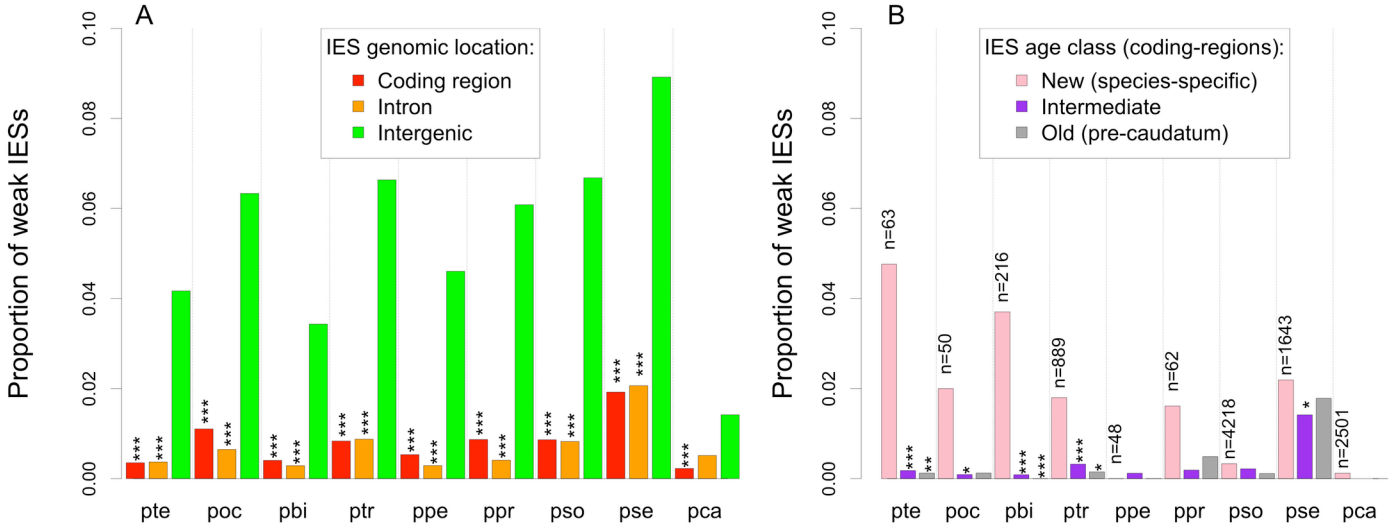

A
Proportion of weak IESs
0.00
0.02
0.04
0.06
0.08
0.10
IES genomic location:
Coding region
Intron
Intergenic
pte
poc
pbi
ptr
ppe
ppr
pso
pse
pca
B
IES age class (coding-regions):
New (species-specific)
Intermediate
Old (pre-caudatum)
n=63
n=50
n=216
n=889
n=48
n=62
n=4218
n=1643
n=2501

Figure S10

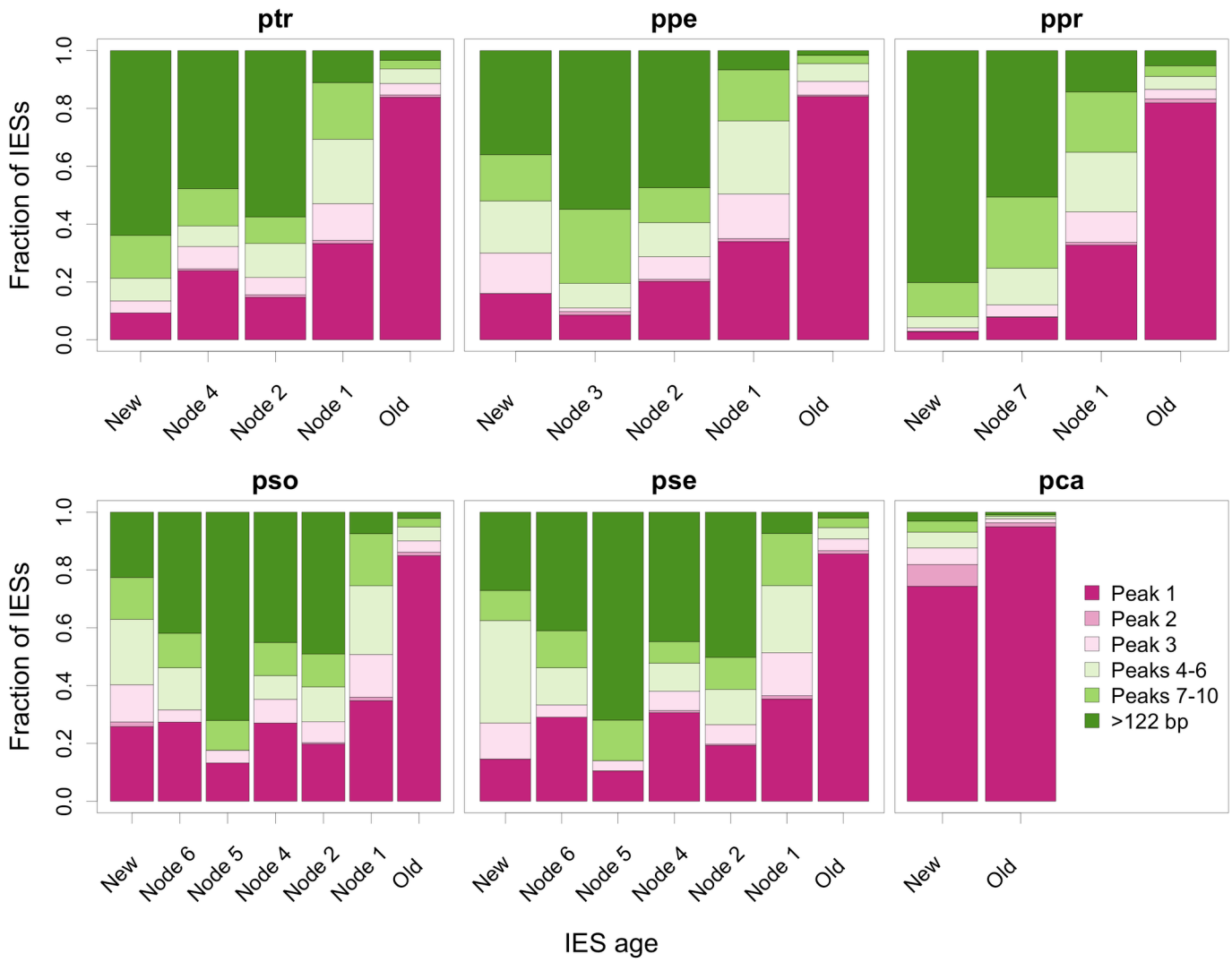

ptr
ppe
ppr
pso
pse
pca
Fraction of IESs
0.0
0.2
0.4
0.6
0.8
1.0
New
Node 4
Node 2
Node 1
Old
Node 3
Node 7
Node 6
Node 5
Peak 1
Peak 2
Peak 3
Peaks 4-6
Peaks 7-10
>122 bp
IES age

Figure S11

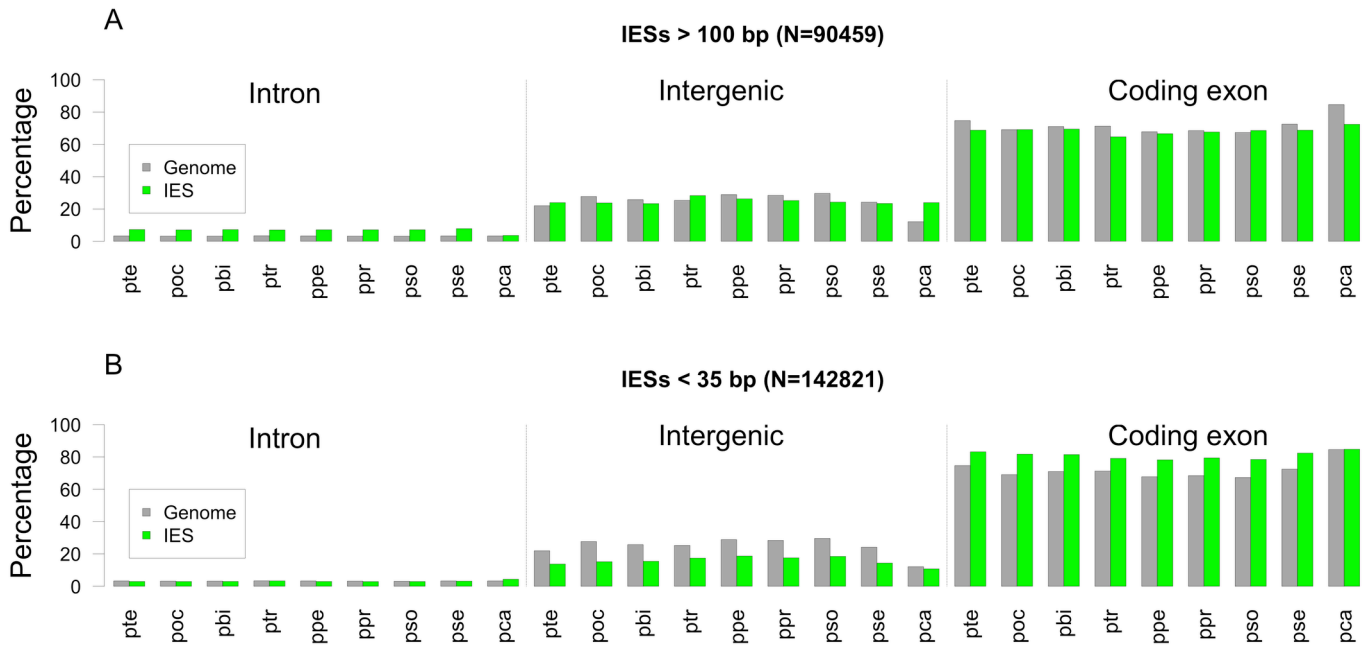

A
IESs > 100 bp (N=90459)
Percentage
100
80
60
40
20
0
Intron
Intergenic
Coding exon
Genome
IES
pte
poc
pbi
ptr
ppe
ppr
pso
pse
pca
B
IESs < 35 bp (N=142821)
Percentage
100
80
60
40
20
0
Intron
Intergenic
Coding exon
Genome
IES
pte
poc
pbi
ptr
ppe
ppr
pso
pse
pca

Figure S12

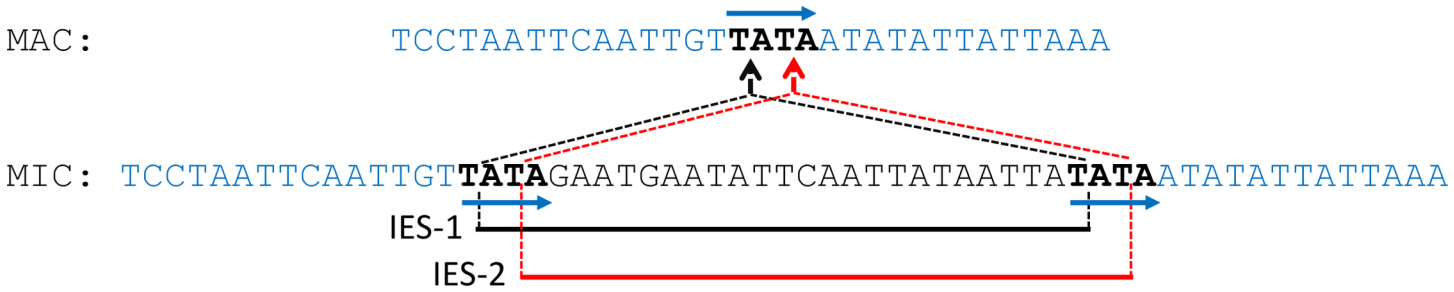

MAC: TCCTAATTCAATTGTTATAATATATTATTAAA
MIC: TCCTAATTCAATTGTTATAGAATGAATATTCAATTATAATTATATAATATATTATTAAA
IES-1
IES-2

Figure S13