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, Institut Jacques Monod, CNRS, 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. Remarkably, we identified 24 families of mobile IESs that generated tens to thousands of new copies. The most active families show the signature of horizontal transfer. These examples illustrate how mobile elements can account for the massive proliferation of IESs in the germline genomes of *Paramecium*, both in non-coding regions and within exons. We also provide evidence that IESs represent a substantial burden for their host, presumably because of excision errors. Interestingly, we observe that IES excision pathways vary according to the age of IESs, and that older IESs tend to be more efficiently excised. This suggests that once fixed in the genome, the presence of IESs imposes a selective pressure on their host, both in *cis* (on the excision signals of each IES) and in *trans* (on the cellular excision machinery), to ensure efficient and precise removal. Finally, we identified 69 IESs that are under strong purifying selection across the *P. aurelia* clade, which indicates that a small fraction of IESs provides a function beneficial for their host. Similar to the evolution of introns in eukaryotes, the colonization of *Paramecium* genes by IESs highlights the major role played by selfish genetic elements in shaping the complexity of genome architecture and gene expression.

Introduction

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

Despite significant progress in characterization of the mechanisms underlying IES elimination, the evolutionary origin of IESs remains mysterious. On the basis of sequence similarities

between the consensus found adjacent to the TA dinucleotide at IES ends and the extremities of DNA transposons from the IS630-Tc1-mariner (ITm) superfamily, Klobutcher and Herrick hypothesized that IESs might be degenerated remnants of transposable elements (TEs)[10,11]. This hypothesis was further substantiated by the discovery that the endonuclease responsible for IES excision in *P. tetraurelia* is encoded by a domesticated PiggyBac transposase [6], assisted by a related family of catalytically inactive transposases [7]. All-by-all sequence comparison of the *P. tetraurelia* 45,000 IESs and of their flanking sequences identified eight families of "mobile IESs" (2 to 6 copies), *i.e.* homologous IESs inserted at non-homologous sites in the genome [4]. One such family (with 6 copies) was found similar to the Terminal Inverted Repeats of *Thon*, a DNA transposon of the ITm superfamily, indicating that some IESs behave as non-autonomous TEs [4]. These cases provided support to the notion that at least some IESs have derived from recently mobilized elements. However, the rather small number of mobile IESs detected (23 copies out of 45,000 IESs) suggested a limited activity of transposable IESs in the recent evolutionary history of the *P. tetraurelia* lineage [4]. There is also evidence that some IESs originated from MAC sequences, as described for instance for the IESs involved in mating type determination in several species [12,13]. The extent to which the 45,000 IESs detected in *P. tetraurelia* derive from TEs or from MAC sequences therefore remained unclear.

In order to gain insight concerning the evolutionary origin of IESs in the *Paramecium* lineage, we adopted a comparative genomic approach. *P. tetraurelia* belongs to the *Paramecium aurelia* group of species that comprises over a dozen morphologically similar yet genetically isolated species [14–17]. Here, we selected eight *P. aurelia* species and one outgroup (*P. caudatum*), and sequenced their germline MIC genomes. Comparison of the IES repertoire across these nine species revealed that IES gains and losses occurred throughout the whole evolutionary history of that clade, with two major waves of insertions: one ancestral wave at the base of the *P. aurelia* clade and one recent wave, specific to the *P. sonneborni* lineage. The analysis of this recent wave revealed thousands of IESs corresponding to mobile elements acquired via horizontal transfer, thus providing the first direct evidence that a majority of IESs can derive from TEs. We also found evidence that IESs represent a substantial burden for their host, because of sub-optimal efficiency of the IES excision process. The comparison of IESs according to their age of origin indicates that over time, IESs shorten and acquire features that allow them to be more efficiently excised. Interestingly, although most IESs diverge very rapidly, we identified 69 IESs that are under strong purifying selection across the *aurelia* clade, which indicates that some IESs provide a function beneficial for their host. The evolutionary history of *Paramecium* IESs is thus reminiscent of the evolutionary history of introns in eukaryotes: selfish mobile elements found a way to invade coding regions and ultimately had a major impact on the biology of the cell and the architecture of its genome.

Results

Sequencing of somatic and germline genomes in nine *Paramecium* species: gigantic germline genome in *P. caudatum*

In order to examine the evolutionary trajectories of IESs in the *Paramecium* lineage, we sequenced the germline MIC genome and the somatic MAC genome of several *Paramecium* species. We selected 8 species from the *Paramecium aurelia* complex and one outgroup species, *P. caudatum*, which diverged from the *aurelia* complex before the two most recent *Paramecium* whole genome duplications [15]. To sequence the germline MIC genome, we purified the germline nuclei (MICs) of each species using a flow cytometry procedure that we

previously developed for *P. tetrauralia* [5]. The strategy consists in a fractionation step to obtain MIC-enriched samples from exponentially growing *Paramecium* vegetative cultures, which are then subjected to sorting by flow cytometry (S1 Fig) (see Materials and Methods). This allows the separation of the small, diploid MICs from the highly polyploid MAC and the bacteria abundant in *Paramecium* cultures. MIC DNA was obtained from highly enriched sorted nuclei (97-99%) for the nine selected *Paramecium* species and was used for paired-end Illumina sequencing (see Materials and Methods and S1 Table).

The MAC genome of the same strains was sequenced as well for four species for which it was not already available (S2 Table). In these four genome assemblies, we observed regions of low coverage at the extremities of MAC scaffolds (S2 Fig). These regions (hereafter referred to as '*MAC-variable*' regions) result from the variability of programmed genome rearrangement patterns during MAC development [18]. While most MIC loci are either fully eliminated during MAC development (MIC-limited sequences) or fully retained (MAC-destined sequences), MAC-variable regions correspond to DNA sequences that are not completely eliminated and instead, are retained in a small fraction of MAC copies. MAC-variable regions represent ~15% of the initial MAC genome assembly (see Materials and Methods and S2 Table). We decided to define the '*constitutive'* MAC genome as the DNA sequences retained in all MAC copies. The size of the constitutive MAC genome assembly was similar among *P. aurelia* species (66-73 Mb) with a noticeably larger size for *P. sonneborni* (83 Mb) (Table 1). The number of protein coding genes follows a similar distribution (36,179 to 42,619) in *aurelia* species, with a larger number of genes (49, 951) in *P. sonneborni* (Table 1). This contrasts with the much smaller MAC genome size (30.4 Mb) and number of genes (18,173) of the outgroup *P. caudatum* [15].

Species (strain)	MIC genome size (Mb)		MAC-destined regions			
	flow cytometry (a)	k-mer (b)	size (Mb) (c)	Nb. of protein genes	Nb. of IESs	IES density (per kb) (d)
P. tetraurelia (51)	151	108 (160)	70	40,460	44,128	0.62
P. octaurelia (138)	175	108	72.6	44,398	44,509	0.61
P. biaurelia (V1-4)	179	119	77	40,261	45,384	0.65
P. tredecaurelia (209)	142	127	66	36,179	42,275	0.66
P. pentaurelia (87)	154	112	72.7	41,676	42,686	0.57
P. primaurelia (AZ9-3)	168	114	73.5	42,619	43,766	0.59
P. sonneborni (ATCC 30995)	458	286 (316)	82.6	49,951	60,198	1.05
P. sexaurelia (AZ8-4)	205	123 (164)	68.0	36,094	47,002	0.70
*P. caudatum (*My43c3d*)*	1,659	1,300	30.5	18,673	(e) 8,762	0.47

Table 1. Characteristics of analyzed genomes.
Species are ordered according to the phylogeny (Fig 1). MIC genome size (in Mb) was estimated based on (a) flow cytometry analysis and (b) k-mer counts. Size estimation before correction, based on MAC contamination, is indicated in parentheses (see Material and Methods). (c) Size of constitutive MAC genome assembly. (d) The IES density was measured in MAC-destined sequences, after exclusion of regions with insufficient MIC read depth (< 15X). (e) The sensitivity of IES detection in *P. caudatum* was limited because of the relatively low sequencing depth of its MIC genome. Based on the IES density observed in regions with sufficient read depth (see c), we estimate that the genome of *P. caudatum* should contain about 15,000 IESs in MAC-destined regions.

To estimate the size of the MIC genomes, we employed two distinct approaches. First, we used the MIC-enriched preparations from *Paramecium* cultures to yield values for DNA quantity in the MICs by flow cytometry. We measured the absolute DNA content in the nuclei with propidium iodide, a fluorophore that is insensitive to differences in base composition, and

compared DNA content of MIC-enriched preparations to a standard (tomato nuclei) of known genome size (see Materials and Methods and S1 Data). The estimated MIC genome sizes are within a similar range (140-173 Mb) for the *P. aurelia* species, except for *P. sonneborni* (Table 1). The genome size of *P. sonneborni* (448 Mb) was estimated to be roughly the double of the others. The second, independent approach for genome size estimations was based on the sequence reads themselves and used the k-mer method described in [19,20]. It assumes that the total number of k-mers (in this case 17-mers) divided by the sequencing depth is a good approximation for genome size (see Materials and Methods). As shown in S2 Table, the estimated MAC genome sizes were in good agreement with the size of the constitutive MAC genome assemblies. The MIC genome sizes estimated by the k-mer method were comprised between 108 Mb to 123 Mb for the *aurelia* species, with a considerably larger MIC genome (283 Mb) for *P. sonneborni* (Table 1). While the values obtained using the flow cytometry method were greater than those with the k-mer method, the estimated MIC genome sizes were within a similar range for both methods (150 Mb with flow cytometry versus 120 Mb with k-mer for *aurelia*, a roughly double size for *P. sonneborni*) (S3 Fig).

With both methods, the estimated MIC genome size of *P. caudatum* strain My43c3d was the largest among the species analyzed (app. 1,300-1,600 Mb). To confirm this observation, we estimated the genome size of other *P. caudatum* strains, performing the same flow cytometry analysis with MIC-enriched preparations for 9 additional strains. We chose 8 strains that belong to the two major clades A and B described in the *caudatum* lineage, as well as another divergent strain [21]. The data confirmed that the MIC genome size in the *caudatum* lineage is far bigger than that in the *aurelia* lineage and revealed great variations of genome size among the different strains (from 1,600 Mb to 5,500 Mb), even within the same clade (S1 Data). To investigate the composition of the gigantic *caudatum* genomes, we searched for the presence of repeats in the MIC sequence reads of strain My43c3d. We identified two major satellite repeats, Sat1 and Sat2 (332 bp and 449 bp, respectively), which represent 42% and 29% respectively of the MIC genome (S4 Fig). To look for the presence of these two satellite repeats in the other *P. caudatum* strains, we performed PCR amplification on whole cell DNA with specific primers for each repeat. Both Sat1 and Sat2 repeats were detected in the *P. caudatum* strains of the clade B, to which the strain My43c3d belongs (S4 Fig). In contrast, these repeats were not amplified in the other *P. caudatum* strains (S4 Fig), indicating that they are not shared by all *P. caudatum* strains and most likely invaded the MIC genome after the divergence between clades A and B.

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

IES repertoire

IESs were identified by comparing MIC sequence reads to the MAC genome assembly (see Materials and Methods; [4,22]). Overall, the number of detected IESs in MAC-destined sequences is similar across *Paramecium* species (~42,000-47,000 IESs), with the exception of *P. sonneborni* (~60,000 IESs) and *P. caudatum* (~9,000 IESs). It should be noted that the sensitivity of IES detection was limited in *P. caudatum*, due the reduced MIC sequencing depth (13X), resulting from the unexpected huge size of the MIC genome. To circumvent this issue, we compared the IES density across species by taking into account only IESs annotated in

regions with at least 15X depth of MIC sequence reads mapped onto the MAC assembled genome (Table 1): in *P. caudatum*, the density of detected IES sites in MAC-destined regions (0.5 IESs per kb) is only slightly lower than in other species (~0.6 IESs per kb). This suggests that the genome of *P. caudatum* probably contains about 15,000 IESs in its MAC-destined regions.

Our approach is designed to identify IESs only if they are present within loci retained in the MAC. Hence, IESs located in MIC-specific regions (e.g., IESs nested within other IESs [23,24]) remain undetected. Interestingly, in four species whose MAC genome was sequenced at very high depth *(P. octaurelia, P. primaurelia, P. pentaurelia* and *P. sonneborni*), the initial MAC genome assemblies included 10 to 16 Mb of MAC-variable regions (see above). We identified many IESs in these regions, at a density (0.4 to 0.5 IESs per kb, S2 Table) nearly as high as in MAC-destined regions (Table 1). This suggests that in addition to IESs located in MAC-destined regions, many other IESs are present within MIC-specific regions.

In all species, the vast majority of IESs in MAC-destined regions (73% to 81%) are located in protein-coding exons and ~5% are located in introns. Overall, there is a slight enrichment of IESs within genes (on average, protein-coding genes represent 78% of MAC genomes, and contain 83% of the IESs; S3 Table). This enrichment is not true for all gene categories. In particular, we observed a depletion of IESs in highly expressed genes: on average, the IES density in the top 10% most expressed genes is 37% lower than in the bottom 10% (S5 Fig). This pattern, consistent with previous observations in *P. tetraurelia*, suggests that IES insertions are counter-selected in highly expressed genes [4].

Age distribution of IESs

In order to explore the origin and evolution of IESs, we resolved the phylogenetic relationship among the sequenced species. To do so, we classified all protein sequences into families (N=13,617 gene families) and inferred the species phylogeny using the subset of 1,061 gene families containing one single sequence from each species. In agreement with previous reports [16,25], we found strong support for a division of the *aurelia* complex in two subclades (hereafter referred to as subclades A and B), separating *P. sonneborni* and *P. sexaurelia* from the other *aurelia* species (Fig 1). We then used this species phylogeny to identify gene duplications and speciation events in each of the 13,617 gene families, using the PHYLDOG tree reconciliation method [26].

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

In order to date events of IES gain or loss, it is necessary to identify IESs that are homologous, i.e. that result from a single ancestral insertion event. For this, we mapped the position of IES

excision sites in multiple alignments of each gene family (nucleic sequence alignments based on protein alignments): IESs located at the exact same position within a codon were assumed to be homologous (S6 Fig). To avoid ambiguities due to low quality alignments, we only analyzed IESs present within well-conserved protein-coding regions (which represent from 45% to 51% of IESs located in coding regions; Fig 1). We then used the reconciled gene tree to map events on the species phylogeny and estimate rates of IES gain and loss along each branch of the species tree using a Bayesian approach accounting for IES losses and missing data (see Materials and Methods). In the absence of fossil records, it is impossible to date speciation events (in million years). We therefore used sequence divergence (number of amino-acid substitutions per site) along branches of the phylogeny as a proxy for time.

Overall, 10.8% of IESs detected in *aurelia* species predate the divergence from *P. caudatum* (referred to as 'Old' IESs in Fig 1), 79% were gained after the divergence of *P. caudatum*, but before the radiation of the *aurelia* complex ('Intermediate' in Fig 1) and 10.2% are more recent. The rate of IES gain varied widely over time: a burst of insertions occurred in the ancestral branch leading to the *aurelia* clade, followed by a progressive slowdown in most lineages, except in *P. sonneborni* where the rate of IES gain strongly increased again in the recent period (18.8% of IESs detected in *P. sonneborni* are specific to that species). The IES gain rate has remained substantial in *P. sexaurelia* and *P. tredecaurelia*, but has dropped to very low levels in *P. tetraurelia/P.octaurelia* and in *P. pentaurelia/P. primaurelia* lineages, about 20 times lower than in *P. sonneborni* or in the ancestral *aurelia* lineage (Fig 1). The rate of IES loss appears to be more uniform along the phylogeny, with only 2 to 3-fold variation (Fig 1).

Recent waves of mobilization of IESs

The episodic bursts of IES gains that we observed in the phylogeny are reminiscent of the dynamics of invasion by TEs. To test the hypothesis that IESs might correspond to TEs, we searched for evidence of mobile IESs, i.e. homologous IES sequences inserted at different (non-homologous) loci. In a first step, we compared all IESs against each other with BLASTN to identify clusters of homologous IESs. In a second step, all clusters with ≥ 10 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 repeated element is located within the IES but does not include the extremities of the IES, and 2) cases where the extremities of the repeated element 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	Number of repeat-containing IESs						Number of mobile IESs per species								
		Total	% CDS	% Intron	% Interg.	Nested repeats	Mobile IESs	pca	pse	pso	ptr	ppe	ppr	pbi	poc	pte
FAM_2183	233	5221	68.9%	7.6%	23.4%	1068	4153	0	4	3252	897	0	0	0	0	0
FAM_3	290	2658	67.6%	8.8%	23.6%	875	1783	0	0	0	15	344	321	766	146	191
FAM_2938	765	1548	68.6%	8.9%	22.5%	1170	378	0	7	370	1	0	0	0	0	0
FAM_2317	768	559	54.2%	5.4%	40.4%	228	331	0	82	140	109	0	0	0	0	0
FAM_2942	211	163	62.6%	6.7%	30.7%	53	110	0	0	110	0	0	0	0	0	0
FAM_2334	214	124	76.6%	7.3%	16.1%	18	106	0	17	89	0	0	0	0	0	0

FAM_2321	471	200	55.0%	3.5%	41.5%	116	84	1	11	58	10	1	1	2	0	0
FAM_78	50	65	0.0%	0.0%	100.0%	9	56	0	0	0	0	34	22	0	0	0
FAM_1402 (TIR *Thon)*	693	109	36.7%	1.8%	61.5%	56	53	0	6	8	5	0	0	3	16	15
FAM_1257 (TIR *Merou)*	522	109	34.9%	7.3%	57.8%	60	49	0	0	5	5	3	2	7	12	15
FAM_670	46	45	73.3%	0.0%	26.7%	1	44	0	0	9	1	2	5	2	24	1
FAM_2649	762	73	47.9%	9.6%	42.5%	33	40	0	0	16	24	0	0	0	0	0
FAM_1473	98	33	72.7%	3.0%	24.2%	0	33	0	0	4	7	1	0	5	10	6
FAM_51	231	75	57.3%	2.7%	40.0%	43	32	0	0	0	0	12	9	11	0	0
FAM_692	93	28	89.3%	3.6%	7.1%	2	26	0	0	4	13	5	4	0	0	0
FAM_1294 (Baudroie)	1706	72	51.4%	4.2%	44.4%	46	26	0	0	0	0	1	2	18	4	1
FAM_2314 (DDE)	3421	480	56.9%	5.8%	37.3%	456	24	0	20	2	2	0	0	0	0	0
FAM_2802	32	23	100.0%	0.0%	0.0%	1	22	0	0	2	20	0	0	0	0	0
FAM_3194	230	26	80.8%	0.0%	19.2%	6	20	20	0	0	0	0	0	0	0	0
FAM_837	50	18	72.2%	5.6%	22.2%	0	18	0	0	0	0	14	4	0	0	0
FAM_1165	77	16	87.5%	0.0%	12.5%	1	15	0	0	0	0	1	1	0	12	1
FAM_2936	223	40	72.5%	10.0%	17.5%	25	15	0	0	15	0	0	0	0	0	0
FAM_1259	231	28	64.3%	0.0%	35.7%	14	14	0	0	0	0	1	0	13	0	0
FAM_3023	350	64	46.9%	3.1%	50.0%	53	11	0	0	11	0	0	0	0	0	0
Total		11777				4334	7443	21	147	4095	1109	419	371	827	224	230

Table 2. Genomic and taxonomic distribution of mobile IESs.
Detected repeats are divided in two categories: nested repeats (i.e. copies inserted within an IES, but not including the extremities of the IES) and mobile IESs (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 mobile IESs in its genome. 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 to *Baudroie*, a composite Tc1-mariner element, and includes an ORF with similarity to tyrosine-type recombinases. FAM_1402 (0.7 kb) and FAM_1257 (0.5 kb) correspond to non-autonomous elements, homologous to the terminal inverted repeats (TIR) of *Thon* and *Merou* respectively. The other families of mobile IESs do not match with any known TEs. Their relatively short lengths (32 bp to 765 bp) and the absence of homology with any known protein, indicate that they most probably correspond to non-autonomous elements, mobilized by transposases expressed from active TEs.

The genomic distribution of mobile IESs within MAC-destined regions is similar to that of other IESs: most of the families are predominantly located within protein-coding regions (which represent ~70% of the MAC genome) (Table 2). The only notable exceptions are FAM_1257, FAM_1402 and FAM_78 elements, which are under-represented within genes (Table 2). In particular, FAM_78 elements are exclusively found in intergenic regions.

As explained previously, it is possible to date insertions for the subset of IESs located within well-conserved protein-coding regions. The vast majority (97.5%) of mobile IES copies that can be dated correspond to recent insertions (as compared to only 9.5% of recent insertions for the other IESs). FAM_3 is present in all genomes of the subclade A (Table 2), and 94% of dated insertions are shared by at least two species, which indicates that this element has been very active at the beginning of the radiation of this clade. For the other families of mobile IESs, more than 97% of insertion loci are species-specific. Thus, all the families of mobile IESs that we detected have been subject to recent waves of insertion. This most probably reflects the fact that more ancient families are difficult to recognize, because of the rapid divergence of IES sequences.

The largest family (FAM_2183) corresponds to a 233 bp-long non-autonomous element, for which we detected a total of 3,252 copies of mobile IESs in the genome of *P. sonneborni*, and 897 in *P. tredecaurelia* (Table 2, Fig 2). Among the 1,973 copies inserted in well-conserved coding regions, only two are shared by the two species. This indicates that this element has been highly active, independently in the *P. sonneborni* and the *P. tredecaurelia* lineages. The very low number of shared copies suggests that these two copies correspond to independent insertion events at a same site, rather than ancestral events. It is important to note that *P. sonneborni* and *P. tredecaurelia* belong to two distantly related subclades of the *aurelia* complex (Fig 1). The high level of sequence similarity between copies (average pairwise identity=72%; Fig 2) and the absence of copies in other *Paramecium* species (except 4 copies in *P. sexaurelia*), indicate that both *P. sonneborni* and *P. tredecaurelia* have been invaded recently by this mobile element. Interestingly, there are four other families (FAM_2317, FAM_2321, FAM_2649, FAM_2802) that are shared by *P. tredecaurelia* and the *P. sonneborni/P. sexaurelia* clade, which implies that multiple families of mobile IESs have been horizontally transferred between those lineages.

Fig 2. Phylogenetic analysis of the largest family of mobile IESs.
(A) Sequence logo [27], based on the alignment of the entire FAM_2183 family (N=4,153 mobile IESs). All copies present a high level of sequence similarity (average pair-wise identity 72%) throughout their entire length (233 bp), not just at their ends. (B) 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.

IES excision mechanism varies with IES age

Like any biological process, the excision of IESs during new MAC development is not 100% efficient [4,29]. For instance, the IES retention rate in *P. tetraurelia* MAC chromosomes is on average 0.8% in wild-type cells [30]. 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 (S7A 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%) (S7B Fig). This indicates that after their insertion, IESs progressively accumulate changes that make them more efficiently excised, presumably in response to the selective pressure against retention of IESs within coding regions.

In *P. tetraurelia*, functional analyses have revealed that different classes of IESs rely on different excision pathways [30–32]. A large subset of IESs (63%) require the histone H3 methyltransferase Ezl1 for their excision, while a much smaller subset (7%) requires both the Ezl1 and the Dcl2/3 proteins, which are necessary for the biogenesis of 25 nt long scnRNAs [30,33]. The remaining 30% of IESs require neither Ezl1 nor Dcl2/3 to complete excision. Using published IES excision efficiency datasets upon silencing of *EZL1* and *DCL2/3* [30], we found that 92% of newly inserted *P. tetraurelia* IESs are sensitive to Ezl1, as compared to 39% for old ones (Fig 3). Similarly, the proportion of Dcl2/3-dependent IESs varies from 17% for new IESs to 3% for old ones. These observations suggest that newly inserted IESs, like TEs themselves [34], initially depend on histone marks deposited by Ezl1 (and to some extent on the scanRNA pathway). Over time, histone marks and scnRNAs become dispensable as IESs gradually acquire features that allow them to be efficiently excised.

Fig 3. Older IESs are less dependent on Ezl1 and Dcl2/Dcl3 for their excision.
Barplots represent the fraction of *P. tetraurelia* IESs with a high retention score (IRS > 10%) after silencing of *EZL1* or *DCL2/DCL3*, according to their age. The age of an IES insertion is defined by the phylogenetic position of the last common ancestor (LCA) of species sharing an IES at the same site (New: *P. tetraurelia*-specific IES; Node *n*: the LCA corresponds to node number *n* in the species phylogeny; Old: the LCA predates the *P. aurelia*/*P. caudatum* divergence).

We also compared the length of IESs according to their age. IESs have a characteristic length distribution, which shows the same ~10 bp periodicity in all *aurelia* species (Fig 4A), likely reflecting structural constraints on the excision process [4,7]. We observed that the length distribution of IESs changes drastically over evolutionary time. For instance, in *P. sonneborni*, *P. tredecaurelia* and *P. tetraurelia*, the proportion of IESs in the first peak of the length distribution (< 35 bp) ranges from 1-10% for new IESs to 81-84% for old ones (Fig 4B), and similar patterns are observed in all other *aurelia* species (S8 Fig). In *P. caudatum*, the overall length distribution is shifted towards shorter IESs (71% in the first peak, compared to 35% in *aurelia*; Fig 4A). This suggests that this lineage has not been subject to IES insertion waves for a long period of time, in agreement with the paucity of recognizable mobile IESs in that genome (Table 2).

Fig 4. Length distribution of IESs according to their age.
(A) Comparison of the length distribution of IESs in *P. caudatum* (N=8,172 IESs) and in species from the *aurelia* clade (N=392,082 IESs). The fraction of IESs present within each peak of the distribution is indicated for the first 10 peaks. (B) Comparison of the length distribution of IESs according to their age (for the subset of datable IESs located in coding regions). The age of IES insertions is defined as in Fig 3. Results from other species are presented in S8 Fig.

Genomic distribution of IESs according to their age

Because of the rapid divergence of non-coding sequences, it is generally not possible to assess homology among IES insertion sites located in intergenic regions, and hence it is not possible to date them directly. We therefore used the length of IESs as a rough proxy for their age, to investigate their genomic distribution over time. We observed that in all *aurelia* species, long IESs (> 100 bp, presumably young) are uniformly distributed across genomic compartments (introns, coding regions and intergenic regions) (S9 Fig). Conversely, short IESs (<35 bp, presumably older) are enriched in coding regions (on average, 81% of short IESs in coding regions, vs. 70% expected; S9 Fig). This suggests that IESs located within intergenic regions have a shorter lifespan than those located in coding regions.

Exaptation of the IES excision machinery

A large majority of detected IESs predate the divergence of the *aurelia* clade (Fig 1). Because of the rapid evolution of non-coding sequences, orthologous IESs from different species are generally too divergent to be recognized by sequence similarity search. Yet the comparison of all sequences against each other revealed several interesting exceptions. Overall, we identified 69 families of homologous IESs conserved across at least 5 of the 8 species of the *aurelia* clade. These highly conserved IESs are similar to other IESs in terms of length (mean=75 bp) or genomic distribution (79% within protein-coding genes, 21% in intergenic regions). Their high levels of sequence conservation indicate that they are subject to strong selective constraints, and hence that they have a function, beneficial for *Paramecium*. By definition, IESs are absent from the MAC genome so they cannot be expressed in vegetative cells. However, they can potentially be transcribed during the early development of the new MAC, before IES excision occurs [31,35]. To gain insight into their possible functions, we analyzed the transcription of conserved IESs using polyadenylated RNAseq data from autogamy time course experiments in *P. tetraurelia* [36]. Among the 56 families of highly conserved IESs present in *P. tetraurelia*, 10 (18%) are transcribed at substantial levels (>1 RPKM) during autogamy (as compared to 0.8% for other IESs) (S4 Table). One of these IESs (~800 bp-long) contains a gene encoding a putative DNA-binding protein, well conserved in all species of the *aurelia* clade and expressed at high levels during the early stages of autogamy (Fig 5).

Fig 5. A highly conserved IES contains a gene encoding a putative DNA-binding protein. (A) Phylogenetic tree of the IES family FAM_4968. This IES is highly conserved in all species of the *aurelia* clade (≥75% nucleotide identity between the most distantly related species). BioNJ tree for 211 sites, Poisson model, 100 replicates. (B) Multiple alignment of the protein encoded by this IES. The coding region is subject to strong purifying selection (dN/dS=0.14). The encoded protein contains a helix-turn-helix DNA-binding domain (PF03221, IPR006600). (C) Gene annotation and expression level during autogamy of *P. tetraurelia*. The IES is located within a gene (on the opposite strand). The gene within the IES is expressed at high levels during the early stages of autogamy (T0 and T5).

All other highly conserved IESs are much shorter (< 300 bp), most probably too short to encode proteins. But we found examples suggesting that some of them contribute to the regulation of the expression of their host gene. For instance, we identified a conserved IES located at the 5' end of a gene of unknown function, encompassing the transcription start site and the beginning of the first exon (including the 5'UTR and the first codons). The excision of the IES during MAC development leads to the loss of the initiation codon and of the promoter region, and thereby to the silencing of this gene in vegetative cells (S10 Fig). These examples illustrate that the IES excision machinery has been recruited during evolution to contribute new functions beneficial for *Paramecium*.

Discussion

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

To explore the evolutionary origin of IESs, we analyzed the MIC genomes of eight species of the *P. aurelia* complex, and of an outgroup species, *P. caudatum*. Unexpectedly, we discovered that the MIC genomes of *P. caudatum* strains are at least one order of magnitude larger than those of *P. aurelia* species (~1,600 to 5,500 Mb vs ~110-160 Mb). The sequencing of *P. caudatum* My43c3d revealed that its huge MIC genome size is caused by the amplification of

two major satellite repeats, which represent 71% of its MIC-limited genome (S4 Fig). The high variability of genome sizes across the *P. caudatum* lineage makes this clade an attractive model system to study the possible phenotypic consequences of genome size variations within a species.

All the *Paramecium* MIC genomes we sequenced present a high density of IESs in MAC-destined sequences: from 0.5 IES per kb in *P. caudatum* up to 1 IES per kb in *P. sonneborni* (Table 1). The vast majority of these IESs (83% on average) are located within genes, as expected given the very high gene density in MAC genomes (S3 Table). In *aurelia* species, there are on average 0.95 IESs per protein-coding gene. The IES density varies among genes, but overall, ~50% of the ~40,000 genes contain at least one IES. Moreover, the analysis of MIC-specific regions that are occasionally retained in the MAC (MAC-variable regions) revealed similar IES densities (S2 Table), which suggests that, in addition to IESs located in MAC-destined regions, many other IESs are located within MIC-specific regions.

To explore the origin and evolution of these tens of thousands of IESs, we sought to identify homologous IESs across the 9 *Paramecium* species. Given their rapid rate of evolution, homologous IESs are generally too divergent to be recognized by sequence similarity at this evolutionary scale. However, it is possible to identify homologous IESs based on their shared position within multiple alignments of homologous genes. Thus, for the subset of IESs located in coding regions, we were able to infer rates of IES gain and loss across the species phylogeny (Fig 1). Overall, about 90% of IESs detected in *aurelia* species predate the radiation of that clade, but fewer than 10% are shared with *P. caudatum*. Thus, the vast majority of *aurelia* IESs result from a major wave of IES gains that occurred after the divergence of *P. caudatum,* but before the radiation of the *aurelia* complex. Similarly, 80% of IESs detected in *P. caudatum* are specific to that lineage, which implies that multiple independent events of massive IES invasions occurred during evolution.

The burst of IES gains at the base of the *aurelia* clade was followed by a progressive slowdown in most species, except in the *P. sonneborni* lineage, which has been subject to a second wave of IES insertions (Fig 1). Interestingly, the comparison of IES sequences revealed that thousands of these insertions result from the recent and massive mobilization of a small number of IESs. Several families of mobile IESs present homology with known ITm transposons, and some of them encode transposases. But most mobile IESs do not appear to have any protein-coding potential, and therefore must correspond to non-autonomous elements, whose mobility depends on the expression of active transposons. The number of detected mobile IES copies varies widely across species (Table 2). For instance, mobile IESs have been very active in the *P. sonneborni* lineage (4,095 copies), much more than in its sister lineage, *P. sexaurelia* (147 copies). Thus, mobile IESs account for at least 20% of the difference in IES number between these two species (Fig 1).

Most IESs found in present-day genomes correspond to unique sequences (S11 Fig). After a burst of transposition, the different copies inserted in the genome are expected to diverge rapidly, like any neutrally evolving sequence. Typically, the average synonymous divergence (measured in orthologous protein-coding genes) between *P. sonneborni* and *P. sexaurelia* is around 0.8 substitutions/site. Thus, in the absence of selective pressure, mobile IES copies that predate this speciation event (and *a fortiori* those that predate the radiation of the *aurelia* complex) are expected to be far too diverged to be recognizable. As a result, mobile IESs that can be detected probably represent only the tip of the iceberg. Overall, we found a strong correlation (R^2=0.86, p=8x10^{-4}) between the number of mobile IES copies detected in each

533 species (Table 2) and the rate of IES gain along corresponding branches of the phylogeny (Fig
534 1), which suggests that most gains result from transposition.
535
536 Interestingly, the five most active families in *P. tredecaurelia* all show the signature of
537 horizontal transfer with the distantly related *P. sonneborni* lineage (Table 2). This is notably
538 the case of the largest family that we identified (FAM_2183: 3,252 and 897 copies in each
539 species, respectively; Fig 2). This pattern is reminiscent of the typical life-cycle of many DNA
540 transposons: when a new element enters a genome, it is initially very active and produces a
541 wave of insertions. Its activity then progressively slows down, largely because defense
542 mechanisms become more efficient in the host genome. In the long-term, DNA transposons
543 escape extinction only if they can occasionally be transmitted to a new host [37]. Thus, the
544 variation of IES insertion rates that we observed in the *Paramecium* phylogeny fits very well
545 with the dynamics of TEs: rare episodes of massive invasions (promoted by horizontal transfer
546 to a new host), followed by progressive slowdown of transposition activity.
547
548 TEs are not the unique source of IES gains. Mutations in MAC-destined regions can generate
549 sequence motifs that are recognized by the IES excision machinery, and thereby create new
550 IESs. There is indeed evidence that cryptic IES signals occasionally trigger the excision of
551 MAC-destined sequences [4,29], and that some IESs originated from MAC-destined sequences
552 [12,13]. However, our results suggest that the vast majority of IESs correspond to
553 unrecognizable fossils of mobile elements – as initially proposed by Klobutcher and Herrick
554 [10,11].
555
556 ## The fitness consequences of IES invasions
557 In all *Paramecium* species, we observed a deficit of IESs in highly expressed genes (S5 Fig).
558 As previously reported in *P. tetraurelia* [4], this pattern most probably reflects selective
559 pressure against IES insertions within genes. Indeed, the IES excision machinery (like any other
560 biological machinery) is not 100% efficient: a small fraction of IES copies are retained in the
561 MAC or subject to imprecise excision [29]. Typically, the average IES retention rate in MAC
562 chromosomes is 0.8% in *P. tetraurelia* [30]. For IESs located within genes, such excision errors
563 are expected to have deleterious consequences on fitness, in particular for genes that have to be
564 expressed at high levels [4]. And indeed, in agreement with this hypothesis of selective pressure
565 against IESs within genes, we observed that the proportion of 'weak' IESs (i.e. IESs with a
566 relatively high retention frequency) is much lower in genes than in intergenic regions (S7A
567 Fig).
568 Despite their selective cost, weakly deleterious IES insertions can eventually become fixed by
569 random genetic drift. Once fixed, the fitness of the organism will depend on its ability to
570 properly excise the IES during MAC development. Over time, selection should favor the
571 accumulation of substitutions that increase the efficiency of IES excision. Indeed, we did
572 observe that the proportion of weak IESs decreases with their age (S7B Fig). Interestingly, older
573 IESs, which are also shorter, are less dependent on the Ezl1 and Dcl2/3 proteins (Fig 3). This
574 suggests that after their insertion, IESs progressively acquire features that make them more
575 efficiently excised, by a pathway that requires neither scanRNAs nor histone marks [30].
576
577 Although most IESs appear to behave as selfish genetic elements, this does not exclude that
578 occasionally, some IESs might confer a benefit for their host. While most IESs diverge very
579 rapidly (as expected for neutrally evolving sequences), we identified 69 families of homologous
580 IESs that have remained strongly conserved across the *aurelia* clade. Their high level of
581 conservation indicates that they are subject to strong purifying selection. This implies that these
582 IESs fulfill a function that contributes to the fitness of *Paramecium*. Notably, we identified one

IES that contains a protein-coding gene (Fig 5). This gene is expressed during the early stages of autogamy, likely from the new developing MAC, before IES excision (Fig 5). Interestingly, 18% of the conserved IESs are transcribed during autogamy (as compared to 0.8% for other IESs). Most conserved IESs are too short to encode proteins, but they may contribute to gene regulation (e.g. S10 Fig). Given the enrichment of conserved IESs in genes expressed during early autogamy, it is tempting to speculate that these IESs may play a role in controlling the IES excision machinery itself. Indeed, this machinery must be tightly regulated to ensure that all IESs are efficiently excised, while limiting off-target excision of MAC-destined regions, which occurs occasionally in MAC chromosomes [4,29]. Thus, developmental disruption of genes encoding IES excision factors by the excision machinery may provide a simple regulatory feedback loop to decrease the activity of the IES excision machinery as soon as a large fraction of IESs have been excised: if a given IES drives the expression of a protein factor that is essential for IES excision, then this process is progressively interrupted by the removal of this IES during MAC development. More generally, such IESs may provide an exquisite developmental process to regulate DNA elimination events and /or MAC differentiation.

Why are IESs not eliminated from the germline genome?

Overall, ~50% of *Paramecium* genes contain at least one IES. Because of excision errors, this high prevalence of IESs within genes must represent a substantial burden. This raises the question of why IESs do not get eliminated from the MIC genome.

In all species, we observed that the length of IESs is negatively correlated with their age (Fig 4, S8 Fig). This pattern is similar to that observed in other eukaryotes, where fixed copies of TEs tend to shrink over time and finally disappear, due to the accumulation of small deletions [38]. IESs located in non-coding regions can be lost by several processes. First, mutations within excision signals (e.g. in the TA dinucleotides) can transform an IES into a MAC-destined sequence. Second, deletions can lead to the loss of an IES – either progressively by successive small deletions or by a single larger deletion encompassing the IES. However, for an IES located within an exon, most deletions affecting the coding-region, and any mutation within the IES preventing its proper excision during MAC development, would be strongly counter-selected. Thus, exonic IES losses can only occur by precise complete deletions that leave the open reading frame intact. We did observe such cases of precise loss (S6 Fig). One possible mechanism is that the IES excision machinery, which is normally at work during MAC development, might occasionally operate within the MIC. An alternative hypothesis is that IESs might be lost from the MIC by gene conversion, through homologous recombination with MAC-derived DNA fragments. Interestingly, this scenario might explain cases where we observed concomitant losses of neighboring IESs (see e.g. IES 5 and 6 in S6 Fig). Further studies will be needed to determine the mechanisms underlying precise IES loss. With regard to the evolution of the number of IESs, the important point is that the rate of IES loss has remained quite stable and relatively low across the phylogeny (Fig 1). Conversely, the rate of IES gains has been much more erratic, characterized by episodic waves of insertions, during which the IES gain rate largely exceeded the loss rate (Fig 1). In the end, the number of IESs reflects the balance between gain and loss rates. Thus, the large number of IESs in *Paramecium* can simply be explained by massive invasions of mobile IESs, followed by periods of lower activity, during which IES copies progressively diverge, and occasionally get lost by deletion from the MIC.

Parallel scenario for the evolution of IESs and spliceosomal introns

In most organisms, gene regulatory elements and coding regions constitute a no man's land for TEs, because insertions that disrupt gene function are strongly counter-selected. But in some ciliates, it is possible for mobile elements to proliferate within genes in the MIC genome, as

long as they are efficiently and precisely excised during the development of the MAC genome, before genes start to be expressed. DNA transposons encode transposases that allow their mobilization by a 'cut-and-paste' process. Generally, the excision step leaves a few nucleotides at the original insertion site, but one peculiarity of PiggyBac transposases is that they can excise copies precisely, without leaving any scar [39]. This feature may have predisposed PiggyBac to extend its niche to genic regions in ciliates. We speculate that the very first proto-IESs corresponded to PiggyBac elements that had evolved a specific transposase with a 'cut and close' activity targeted to the developing MAC. As soon as several copies of these proto-IESs have been fixed within genes, then the host organism has become dependent on the activity of the PiggyBac transposase to ensure that all these copies are precisely excised from its MAC. This selective pressure would have driven the domestication of the PiggyBac transposase by its host, and then, progressively, the evolution of the other components that contribute to the efficient excision of proto-IESs. Once the IES excision machinery is in place in the ancestral *Paramecium* lineage, other families of TEs (including non-autonomous elements) could hijack the machinery and in turn exploit this intragenic niche, eventually creating the tens of thousands of IESs found in present-day *Paramecium* genes. The first steps of this scenario remain speculative, since there are no recognizable traces of PiggyBac-related IESs in present-day genomes. But, the discovery of thousands of mobile IESs directly demonstrates the major contribution of TEs to the expansion of the IES repertoire.

This scenario is in many points similar to the one proposed for the evolution of spliceosomal introns. Indeed, it had long been postulated, based on similarities in biochemical processes, that spliceosomal introns derive from mobile elements (group II self-splicing introns) [40]. In eukaryotes, the spread of introns in protein-coding genes has been facilitated by the fact that transcription and translation occur in separate compartments, thus offering the opportunity for these mobile elements to be excised from the mRNA in the nucleus without interfering with its translation in the cytoplasm [40] – like IESs, which are excised from genes before they get expressed in the MAC. Once the first introns were established, selection drove the emergence of host factors contributing to the efficiency of the splicing process, which progressively led to the evolution of the modern spliceosome - a complex ribonucleoprotein machinery composed of more than 200 proteins and five small RNAs [41]. In turn, the existence of the spliceosome released the requirement for introns to maintain their self-splicing activity [42], and allowed other TEs to hijack this machinery. The recent discovery of non-autonomous DNA transposons that generated thousands of introns in genomes of some algae directly demonstrated that mobile elements are a major source of new introns [43]. During evolution, the spliceosome has been exapted to fulfill functions useful for the host, notably via the process of alternative splicing, which contributed to diversification of the protein repertoire [44]. Alternative splicing has also been recruited as a means to regulate gene expression [45]. In particular, this is the case of many genes that encode splicing factors, which contain highly conserved introns, allowing them to control the homeostasis of the spliceosome via auto-regulatory loops [46,47]. This pattern is reminiscent of highly conserved IESs that we uncovered in *Paramecium lineages*, which appear to be particularly enriched within genes that are expressed during early MAC development. But although it is clear that some introns have a function, it should not be forgotten that, like IESs, introns also represent a burden for their host, because of errors of the splicing machinery [48–52].

The coexistence of MAC and MIC is a common feature of all ciliates, yet they do not all contain such a high density of IESs in coding regions. Notably, there are ~12,000 IESs in the germline genome of *Tetrahymena thermophila* (~0.1 IES per kb of MAC-destined sequence), but only 11 of them are located within coding regions [53]. These exonic IESs differ from other IESs by

their strongly conserved terminal inverted repeats ending with 5'-TTAA-3', the target site of piggyBac transposons. They are excised precisely (restoring a single TTAA) by two domesticated piggyBac transposases, Tpb1 and Tpb6, which may thus have retained the cleavage specificity of their transposon ancestor [54,55]. We analyzed these 11 exonic IESs: 8 of them are inserted in protein-coding regions that are not conserved in *Paramecium*, and the other 3 are inserted at sites that do not contain IESs in *Paramecium*. There is therefore no evidence for shared exonic IESs between *T. thermophila* and *Paramecium*. The vast majority of the ~12,000 *T. thermophila* IESs are excised by another domesticated piggyBac transposase, Tpb2 [56]. Although Tpb2 retains the cleavage geometry of piggyBac transposases, producing staggered double-strand breaks with 4-nt 5' overhangs [56], it has lost almost all sequence specificity and is thought to be recruited at IES ends by chromatin marks [57]. As a result, several possible cleavage sites are usually present at IES ends and the rejoining of flanking sequences generates microheterogeneity in the MAC sequence [53], which explains why Tpb2-dependent IESs are restricted to introns and intergenic regions [53]. It is important to note that Tpb2 is an essential gene in *T. thermophila* [56], suggesting that genome-wide retention of IESs in the MAC is still highly detrimental. Interestingly, phylogenetic analyses indicate that the *Paramecium* endonuclease PiggyMac (Pgm) and Tpb2 are more closely related to each other than to Tpb1 or Tpb6, and may even be orthologs [7]. In the case of Pgm, however, sequence specificity was relaxed only for the two distal positions of the 4-nt cleavage sites, and the central TAs remain a strict requirement for IES excision in *Paramecium*. Although piggyBac transposons are completely absent from the present-day *Paramecium* germline, this evolutionary solution may have been favored because it also allowed for precise excision of Tc1/mariner insertions, which in turn would have allowed continuous accumulation of insertions within exons [4].

Importantly, the fact that a mechanism of precise excision exists in *T. thermophila* (via Tpb1 and Tpb6) raises the question of why intragenic IESs are not more abundant in its genome. A similar question arises from the distribution of introns in eukaryotes: why are introns very abundant in some lineages but rare in others (e.g. ~7 introns per gene in vertebrates vs ~0.04 in hemiascomycetous yeast)? Part of the explanation may reside in the fact that, because of population genetic forces, some lineages are more subject to random genetic drift than others, and therefore are more permissive to invasion by weakly deleterious genetic elements [51,52]. And it is also possible that the abundance of intragenomic parasites is strongly affected by contingency – rare events of massive invasion, followed by long periods during which copies are lost at a slow rate.

In conclusion, the evolution of the nuclear envelope opened the way for introns to invade genes in eukaryotes, and likewise, the separation of somatic and germline functions between the MIC and the MAC offered the possibility for selfish genetic elements to invade genes in ciliates. Genetic conflicts between these selfish elements and their host genome resulted in the evolution of complex cellular machineries (the spliceosome, the IES excision machinery), which, in the short term, reduced excision errors, but in the long term facilitated their proliferation within genes. The paradigm of intragenomic parasites [58–60] provides a simple and powerful explanation for the "raison d'être" of these mysterious pieces of non-coding DNA that interrupt genes.

Materials and Methods

Cells and cultivation

All experiments were carried out with the *Paramecium* strains listed in Table 1. *Paramecium aurelia* cells were grown in a wheat grass powder (WGP, Pines International, USA) infusion

733 medium bacterized the day before use with *Klebsiella pneumoniae* and supplemented with 0.8
734 mg/L of β-sitosterol (Merck). Cultivation and autogamy were carried out at 27 °C. Monoclonal
735 cultures of the *P. caudatum* cells were grown in a 0.25% Cerophyl infusion inoculated with
736 *Enterobacter aerogenes* at 22°C [61].
737
738 ## Micronucleus-enriched preparation
739 To purify the MICs from vegetative cells, we used the same strategy as the one previously
740 published [5,20], with some optimization for the sorting steps. For *Paramecium aurelia*,
741 transgenic cells expressing a micronuclear (MIC)-localized version of the Green Fluorescent
742 Protein (GFP) were obtained by microinjection of the vegetative macronucleus with the *P.*
743 *tetraurelia* CenH3a-GFP plasmid, described in [62]. In the transformed clones, GFP was
744 exclusively found in the MICs and the transformed clones were selected for their GFP
745 signal/noise ratio. Viability of the sexual progeny after autogamy of the transformed clones was
746 systematically monitored to make sure that the presence of the transgene did not impair the
747 functionality of the MICs. A MIC-enriched preparation was obtained from approximately 3 L
748 of exponentially growing vegetative cells after fractionation and Percoll density gradient
749 centrifugation as described in [5] and kept at -80°C until further use.
750 A slightly different procedure was used for *Paramecium caudatum* cells, which were not
751 transformed with the CenH3a-GFP transgene. The MICs of *P. caudatum* strain My43c3d (used
752 for genome sequencing) were purified with a protocol modified from [63]. Briefly, 3L of a
753 starved culture (~600 cells/mL) were filtered through 8 layers of gauze and concentrated by
754 centrifugation in pear-shaped centrifuge tubes. Packed cells were transferred to a 250 mL cell
755 culture flask, resuspended in 150 mL sterile Eau de Volvic and incubated over night at 22°C.
756 All subsequent steps were performed at 4°C or on ice. The overnight culture was again
757 concentrated by centrifugation and the cell pellet was resuspended and washed in 0.25 M TSCM
758 buffer (10 mM Tris-HCl, pH 6.8, 0.25 M sucrose, 3 mM CaCl2, 8mM MgCl2) [64]. After
759 centrifugation for 3 min at 100 g, pelleted cells were resuspended and incubated for 5 min in
760 10 mL 0.25M sucrose-lysis buffer (10 mM Tris-HCl, pH 6.8, 0.25 M sucrose, 3 mM CaCl2,
761 1mM MgCl2, 0.1% Nonidet-P40, 0.1% Na-deoxycholate). The cell suspension was centrifuged
762 for 2 min at 500 g and the packed cells were lysed in 1 mL of 0.25M sucrose-lysis buffer by
763 about 10-20 strokes on a vortex machine. Lysed cells were washed in 14 mL of 0.25 M TSCM
764 buffer and centrifuged for 1 min at 100 g. The supernatant (containing the MICs) was
765 centrifuged for 10 min at 1,500 g and the pellet was resuspended in 8 mL of 60% Percoll. This
766 suspension was centrifuged for 15 min at 24,000 g in a fixed-angle rotor and the micronuclei
767 formed a diffuse band near the middle of the centrifuge tube. This MIC containing layer was
768 carefully removed with a pipette in about 2 mL, diluted with 10 mL of 0.25 M TSCM buffer
769 and pelleted by centrifugation for 10 min at 1,500 g. The MIC pellet was resuspended in 100
770 µL of 0.25 M TSCM buffer, carefully mixed with 50 µL of 50% glycerol and kept at -80°C
771 until further use.
772 The MICs of the other *P. caudatum* strains were purified with a similar protocol, but omitting
773 the Percoll step and replacing it with centrifugation across a sucrose cushion. Lysed cells were
774 resuspended in 9 mL of 0.25 M TSCM buffer and this suspension was carefully layered on top
775 of a sucrose cushion consisting of 2 mL of 1.6 M TSCM buffer and 2 mL of 0.9 M TSCM
776 buffer and centrifuged in a swinging bucket rotor for 10 min at 300 g with lowest acceleration
777 and braking levels. Depending on the strain, the micronuclei accumulated at the bottom of the
778 0.25 M or 0.9 M TSCM cushion and were removed by careful pipetting of the respective phases
779 to new 15-mL tubes. MIC- containing suspensions were diluted with 0.25 M TSCM buffer,
780 centrifuged for 10 min at 1,500 g and the MIC pellets were subsequently treated as described
781 above.
782

Quantification of MIC DNA content by flow cytometry

MIC-enriched samples were thawed on ice, diluted 1/5 to 1/10 in washing buffer (0.25 M sucrose; 10 mM Tris pH 7.4; 5 mM MgCl2; 15 mM NaCl; 60 mM KCl; 0.5 mM EGTA) and stained on ice with propidium iodide at 100 µg/mL final concentration. We used Tomato nuclei obtained from Montfavet 63-5 hybrid F1 seeds as internal standards of known genome size. Tomato nuclei were obtained from 1 cm^2 of young leaves chopped in a Petri dish with a scalpel. 800 µL of a modified Galbraith buffer [65], containing 45 mM MgCl2, 30 mM Sodium-Citrate and 20 mM MOPS pH 7.0, 40 µg/mL RNAse A, 0.1% Triton X-100, 5 mM sodium metabisulfite (S2O5Na2) was added. The nuclei were collected by pipetting, filtered on 70 µm mesh, and stained on ice with propidium iodide at 100 µg/mL final concentration.
The samples were analyzed on a CyanADP Cytomation analyzer from Beckman-Coulter equipped with 3 lasers: 405 nm, 488 nm and 635 nm. Fluorescence intensity (PE signal in pulse-height) of the nuclei was measured at 575/25 nm, after excitation with the 488 nm laser. Results are deduced from 2C nuclei in individuals considered diploid and are given as C-values [66]. The ratio of fluorescence intensity of 2C-nuclei from sample and standard allows calculation of genome size. C corresponds to the nuclear genome size (the whole chromosome complement with chromosome number n), 1C and 2C being, respectively, the DNA contents of the haploid (n) and diploid (2n) sets of chromosomes. The haploid nuclear DNA content is expressed in picograms or million base pairs, where 1 pg = 978 Mbp [67], considering Tomato 2C DNA (pg) =1.99, according to [68]. The raw data and calculations are provided in S1 Data.

Micronucleus sorting by flow cytometry and flow imaging

To sort the MICs, the MIC-enriched samples were submitted to flow cytometry. *P. aurelia* MICs were sorted based on the SSC, FSC, DAPI (DNA staining), and GFP signals. *P. caudatum* MICs, which are bigger than *aurelia* MICs, could be sorted based on their SSC, FSC, and DAPI signals, without the use of a MIC-specific GFP fluorophore. Quality control was performed by flow cell imaging, using the ImageStreamX (Amnis/Merck Millipore) imaging flow cytometer, as previously described [5]. The MICs represented >99% of the sorted sample, except for *P. sonneborni* (97%). An example of sorting is shown in Fig S1.

Genomic DNA extraction and sequencing

For MAC DNA sequencing, genomic DNA was extracted from vegetative *Paramecium* cell culture after centrifugation and washes with Tris 10mM pH 7.4. For MIC DNA sequencing, DNA was extracted from the sorted MIC samples. The cell or nuclei pellet was treated with 3 volumes of proteinase K solution (0.5 M EDTA pH 9; 1% N-lauroylsarcosine; 1% SDS; 1 mg/mL proteinase K) at 55 °C overnight. Genomic DNA was extracted with Tris–HCl-phenol pH 8 with gentle agitation followed by dialysis against TE (10 mM Tris–HCl; 1 mM EDTA, pH 8) 25% ethanol then against Tris 1 mM pH 8. An RNAse A treatment was performed on MAC DNA, followed by phenol extraction and dialysis as described above. DNA concentration was quantified using QuBit High sensibility kit (Invitrogen) and stored at 4 °C.
As the amounts of DNA extracted from the MIC are too low (30-50 ng), only an overlapping paired-end library could be prepared for *de novo* sequencing. Briefly, 30-50 ng of MIC DNA were sonicated using the E210 Covaris instrument (Covaris, Inc., USA) in order to generate fragments mostly around 500bp. Illumina libraries were then prepared using the NEBNext DNA Sample Prep Master Mix Set (New England Biolabs, MA, USA) and DNA fragments were PCR-amplified using Platinum Pfx DNA polymerase (Invitrogen) and P5 and P7 primers. Amplified library fragments of roughly 500 – 600 bp were size selected on 2% agarose gel. Libraries traces were validated on a Agilent 2100 Bioanalyzer (Agilent Technologies, USA) and quantified by qPCR using the KAPA Library Quantification Kit (KapaBiosystems) on a MxPro instrument (Agilent Technologies, USA). The libraries were sequenced using 251 base-

length read chemistry in a paired-end flow cell on the Illumina HiSeq2500 sequencer (Illumina, USA) in order to obtain overlapping reads that could be fused to generate longer reads of 400-450 bp.

For the MAC genomes, an overlapping paired-end library as described above and four additional mate-pair libraries (about 5Kb, 8Kb, 11Kb and 13Kb) were prepared following Nextera protocol (Nextera Mate Pair sample preparation kit, Illumina). Each library was sequenced using 100 base-length read chemistry on a paired-end flow cell on the Illumina HiSeq2000 (Illumina, USA).

Information about the sequencing data generated for this study is available in S5 Table.

RNA extraction and sequencing

For the purpose of gene annotation, we sequenced mRNAs from vegetative cells (S5 Table). 400 mL cultures of exponentially growing cells at 1000 cells/mL were centrifuged and flash-frozen in liquid nitrogen prior to TRIzol (Invitrogen) treatment, modified by the addition of glass beads for the initial lysis step.

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

MAC Genome assembly

The MAC genomes sequenced for this project were all assembled according to the following steps.

First, long Illumina reads were obtained from 250 bp overlapping paired-end reads sequenced from ~450 bp fragments. The reads were fused with fastx_mergepairs, an in-house tool developed at Genoscope using the fastx library (http://hannonlab.cshl.edu/fastx_toolkit/). An alignment of at least 15 bp with at least 90% identity and fewer than 4 errors was required to fuse two reads into one longer read. The set of fused reads, completed with any reads that could not be fused, was assembled into contigs by the Newbler version 2.9 overlap-layout-consensus assembler, with a minimal alignment identity of 99% and a minimal alignment size of 99 bp. Scaffolds were built from the contigs using 4 Illumina mate-pair libraries with respective insert sizes of 5 kb, 8 kb, 11 kb and 13kb. The scaffolder SSpace [69] was used, with default parameters and an acceptable variation in mate-pair insert size of 25%. Gap closing was a two step process with SOAPdenovo2 GapCloser software [70]. The first step used the Illumina paired-end reads, the second step used the Illumina mate-pair libraries. Finally, Kraken software [71] and the NR nucleotide database were used to detect and remove non-eukaryotic scaffolds, owing mainly to bacterial contaminants (see below).

Filtering

Scaffolds with a length inferior to 2kb or with a G+C content greater than 40% were filtered. Contaminant scaffolds were identified and removed from the assembly provided the Kraken

Kmer score was superior to 10 or a BLASTN match (-evalue 1e-40 –perc_identity 70) against RefSeq database (excluding *Paramecium* sequences) covered at least 20% of the scaffold length. If the mitochondrial genome (more or less fragmented) could be identified by a BLASTN (-evalue 1e-1 –perc_identity 70) against the *P. tetraurelia* mitochondrial genome, the scaffold(s) were tagged as mitochondrial. A handful of chimeric scaffolds were detected and corrected in the *P. octaurelia*, *P. primaurelia* and *P. sexaurelia* assemblies by visual inspection of available long-range sequencing data (remapped mate-pairs) (see S2 Data)).

The constitutive MAC

Paired-end MAC DNA sequencing data were mapped on the MAC genome assembly using Bowtie2 (v.2.2.3 –local, otherwise default parameters) [72]. We defined the constitutive MAC as consisting of all regions of the assemblies with the expected average read coverage. We defined the regions of low coverage at scaffold extremities as MAC-variable regions. In relation to the MAC DNA-seq depth, a minimum expected coverage (v1.9 of samtools depth –q 10 –Q 10) was defined for each assembly (*P. octaurelia* 50X, *P. pentaurelia* 35X, *P. primaurelia* 20X, and *P. sonneborni* 35X). For each scaffold extremity, a Perl script analyzed the coverage in sliding 2kb windows. The first window from the end of the scaffold with a coverage above the minimum expected coverage delimitated the end of the MAC-variable regions. Only regions of minimum size 4 kb were kept. The script adjusted region ends using the MAC telomerisation sites and the ends of coding genes. After this automatic pipeline, each scaffold and mask was adjusted by eye using Circos drawings [73] (see example S2 Fig, representing DNA and RNA coverage, density in non-coding genes and positions of the MAC telomerisation sites). The positions of the regions used to reconstruct the constitutive MAC for each MAC assembly are provided in S2 Data.

IES annotation

Annotation of IESs was performed using the ParTIES toolkit [22] with default parameters. Briefly, this involves (i) alignment of MIC paired-end reads with a reference MAC genome to establish a catalog of potential IES insertion sites and to exclude reads that match perfectly across these sites hence do not contain IESs; (ii) assembly of the remaining reads with Velvet to obtain contigs that may contain IESs; (iii) alignment of the contigs with the MAC reference genome to determine the position and the sequence of the IESs.

Gene annotation

Gene annotation for the 9 species was carried out using a pipeline specifically tuned for the high gene density and tiny intron size (20 – 30 nt) characteristic of *Paramecium* somatic genomes. RNA-Seq transcriptome data was used to predict transcription units with the TrUC v1.0 software (https://github.com/oarnaiz/TrUC), as detailed in [36]. EuGene v4.1 software [74] configured with curated *Paramecium tetraurelia* genes [36] was used for *ab initio* predictions and to combine annotation evidence (the transcription units, the *ab initio* predictions and comparative genomics evidence).

Assembly-free Genome Size Estimation

Illumina paired-end sequencing reads were used to estimate genome size based on counting all substrings of 17 nt in the reads, using jellyfish software version 2.2.10 [75]:

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

The method for genome size estimation, described in [19,20], assumes that the total number of *k*-mers (in this case 17-mers) divided by the sequencing depth is a good approximation for genome size.

As discussed in [20] for *Paramecium* genomes, the histogram of *k*-mer depth for a perfect, homozygous genome with no repeated sequences (and no sequencing errors) is fit by a Poisson distribution, the peak corresponding to sequencing depth. For real genomes, the estimate of genome size is obtained by dividing the total *k*-mer count (excluding the peak near the origin that results from *k*-mers with sequencing errors) by the sequencing depth. This is straightforward for MAC genomes. For MIC genomes, variable amounts of contamination from MAC DNA lead to a second peak at higher *k*-mer depth corresponding to the sum of MAC-destined k-mers in the MIC DNA and the MAC *k*-mers in the contaminating MAC DNA. This was only a significant problem for the *P. tetraurelia*, *P. sexaurelia* and *P. sonneborni* MIC DNA samples, which were approximately corrected by assuming that a proportion of the *k*-mers counted from this second peak up to a depth of 500 were contributed by the contaminating MAC reads, while all the *k*-mers with a depth greater than 500, corresponding to highly repeated sequences, are of MIC origin (S12 Fig). The proportion of contaminating MAC DNA needed for this calculation was confirmed using IES retention scores (IRS) calculated with the MIC sequencing reads [22]. The position of the peak in the IRS distribution indicates the proportion of MIC (IRS ~ 1) and MAC (IRS ~ 0) DNA in the sample, as illustrated in S12 Fig.

Identification of gene families

We performed an all against all BLASTP (ncbi-blast+ v. 2.2.30+) [76] search using the predicted protein sequences from each genome including also the proteins of *Tetrahymena thermophila* (June 2014 assembly http://ciliate.org) as an outgroup. From the resulting output we determined gene families with SiLiX v. 1.2.9 [77]. The resulting gene families were aligned with MAFFT v7.305b (2016/Aug/16) [78] using the --auto option. Gene families with less than 3 genes or average pairwise identity less than 50% where excluded from downstream analyses. From the protein alignments we reconstructed the nucleotide coding sequence alignments.

Paramecium species phylogeny

To reconstruct the species phylogeny, we first selected single-copy gene families present in all nine *Paramecium* species (N=1,061 genes). When available, the *T. thermophila* homolog was also included as an outgroup. We estimated the maximum likelihood phylogeny using IQtree v.1.4.2 [79], considering each gene as a separate partition. We performed model testing on each partition and chose the best codon model (determined by the largest BIC). We evaluated the results by 1,000 bootstrap replicates. All internal branches but one are supported by 100% bootstrap values (Fig 1). We will hereafter refer to this species tree inferred from single-copy gene families as *Tree1*.

The rationale for analyzing single-copy gene families is that these sets of homologous sequences are *a priori* expected to correspond to orthologs. However, given that paramecia have been subject to three rounds of whole genome duplications followed by massive gene losses [80], it is possible that some single-copy gene families include paralogs. To check whether hidden paralogs might have biased the estimation of the species tree, we used PHYLDOG v.2.0beta (build 10/10/2016), a maximum likelihood method to jointly infer rooted species and gene trees, accounting for gene duplications and losses [26]. The analysis was performed using all gene families (N=13,617). The default program options were used with additionally setting a random starting species tree and BIONJ starting gene trees. The duplication and loss parameters were optimized with the average then branchwise option and the genomes were not assumed to have the same number of genes. We also ran PHYLDOG

considering *Tree1* as the fixed species tree, and keeping the remaining options identical. The topology of the most likely species tree inferred with PHYLDOG is almost identical to *Tree1* (it only slightly differs in the positions of *P. biaurelia* and *P. tredecaurelia*), and its likelihood is not significantly different from that obtained when running PHYLDOG with *Tree1* as a species tree. Thus, the species tree inferred by PHYLDOG using all gene families (N=13,617) shows no significant disagreement with the phylogeny based on single-copy gene families (*Tree1*). We therefore hereafter considered *Tree1* as the reference species tree for all our analyses. To identify duplication and speciation nodes in gene phylogenies, we computed reconciled trees for each gene family with PHYLDOG, using *Tree1* as a species tree.

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

To identify events of IES gain and loss along the species phylogeny, it is necessary to analyze the pattern of presence/absence of IESs at homologous loci across species. One difficulty is that some IESs may remain undetected (false negatives). In particular, the sensitivity of ParTIEs depends on the local read coverage [22]. To take into account the uncertainty arising from the variable local read coverage along scaffolds of each species we calculated the coverage of MIC reads mapped against the MAC genome. We identified genes with extreme values of coverage (less than the 10th percentile or more than the 90th over all genes) or with an absolute read coverage of less than 15 reads. These genes correspond to regions with possible assembly errors or to regions of low power to detect IESs, and we marked them as problematic for IES annotation. IESs in these genes were considered to have an uncertain status of presence, and if no IES was annotated the genes were marked as potentially containing IESs. To avoid issues due to genome assembly errors, we excluded from our analyses all IESs identified on small scaffolds (< 10 kb)

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

To identify homologous IES loci, i.e. that result from a single ancestral insertion event, we searched for IESs located at a same site across homologous sequences. It should be noted that the location of IESs, inferred from the comparison of MIC and MAC sequences, is sometimes ambiguous. This occurs when the IES boundaries overlap a motif repeated in tandem (S13 Fig). Such cases, hereafter called "floating IESs", represent 7% of all IESs. In the vast majority of cases (86%) the alternative locations of floating IESs differ by only two bp (as in the example shown in S13 Fig), and there are less than 1% of floating IESs for which the uncertainty in IES position exceeds 5 bp. To determine the exact location of IESs and capture the inherent ambiguity due to possible floating IESs we used a 10bp window around each annotated IES location to determine if the IES was classified as floating. If so, the alternative locations were added to the IES annotation.

Homologous IES insertion sites

To detect homologous IES loci, we compared the position of IESs within homologous genes. To do so, we analyzed gene families with more than 3 sequences and average pairwise identity (at the protein sequence level) of more than 50%. To avoid ambiguity in the identification of homologous sites, we filtered protein sequence alignments with GBlocks v0.91b [81] and we only retained IESs located within conserved alignment blocks. An IES insertion site spans two nucleotides (TA). In a multiple sequence alignment including gaps, an IES locus can be larger (e.g. T--A). Two IES loci were considered as homologous if they have at least one shared site within the alignment (taking into account all potential locations in the case of floating IESs). In the case of floating IESs overlapping the boundaries of conserved alignment blocks, the presence or absence of homologous IES loci in other sequences cannot be reliably inferred. We

therefore only retained IESs for which all homologs (if any) are entirely located within the conserved alignment blocks (i.e. we discarded sets of homologous IES loci that included some floating IESs for which some of the possible alternative positions were located outside of the conserved alignment blocks).

Ancestral state reconstruction and inference of IES insertion and loss rates

To explore the dynamics of IES gain and loss we used a Bayesian approach to reconstruct the ancestral states of presence and absence of IESs using revBayes 1.0.0 beta 3 (2015-10-02) [82]. We constructed binary character matrices (presence/absence) for each gene family containing at least one IES unambiguously located within a conserved alignment block (see above). We assumed a model of character evolution with one rate of gain and one rate of loss sampled from the same exponential distribution with parameter α and a hyperprior sampled from an exponential with parameter 1. We excluded from the analysis 5 gene families for which revBayes could not compute a starting probability due to very small numbers. We used PHYLDOG reconciled gene trees (see above) to fix gene tree topologies and branch lengths. We ran $5x10^5$ iterations. The search parameters were optimized in an initial phase of 10,000 iterations with tuning interval 1,000. Good sampling of the parameter space was verified by inspecting the time series and autocorrelation plots of the parameters. The convergence was validated by inspecting the multivariate Gelman and Rubin's diagnostic plots for different iterations.

Thus, for a given IES locus in a given gene family, revBayes provides an estimate of the probability of presence of an IES at each node of the gene phylogeny. We used these probabilities of presence along the gene phylogeny to estimate rates of IES gains or losses in each branch of the species tree. Because of gene duplications, a given branch in the species tree can be represented by several paths in the gene tree. Thus, we considered all paths in the gene tree that connect the corresponding speciation nodes (see S14 Fig for a simplified example). To measure the IES gain rate at a given IES locus (*c*), in a given gene family (*g*), we define p^{+}_{cgij} as the sum of increase in probability of presence of an IES at this locus along all paths of gene family *g* connecting speciation nodes *i* and *j* (where *i* is a direct ancestor of *j*). Let *ng* be the length in kilobase pairs of gene family *g* alignment (counting only well aligned sites, where the presence of IESs can be assessed). Let *Ig* be the number of IES loci in family *g*. Let *kgij* be the number of paths connecting speciation nodes *i* and *j* in family *g*. Let *bij* be the branch length connecting nodes *i* and *j* in the species tree (*bij* is taken here as a proxy for time). Let p^{+}_{gij} be the sum of increase in probability of presence of an IES, cumulated over all IES loci in family *g*. We define p^{+}_{ij} as the sum of increase in probability of presence of an IES, cumulated over all IES loci along the path *i* to *j* of family *g*. We define *Gij* as the rate of IES gain over all gene families (*f*) along path *ij* expressed in number of IES gains per kilobase pairs of alignment per unit of time:

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

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

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

IES age of insertion

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

Identification of homologous IES sequences and characterization of mobile IESs

To characterize families of homologous IES sequences, we first compared all IESs (from all species) against each other with *blastn* (ncbi blast+ v2.5.0, [76]):

```
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 [77] 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 [83].

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.4415828).

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.

Data availability

All sequences and genome assemblies have been deposited in public databases (accession numbers in S5 Table). All detected IESs and their annotation have been deposited at (https://doi.org/10.5281/zenodo.4415828). This archive also contains the list of mobile IESs and their alignments, the list of highly conserved IESs and their alignments.
All scripts used in the analysis are available at https://github.com/sellisd/IES

Funding

This work was supported by the Centre National de la Recherche Scientifique, by the Agence Nationale de la Recherche (ANR-18-CE12-0005-04; ANR-19-CE12-0015-01), and by the Fondation de la Recherche Medicale (Equipe FRM DEQ20160334868) to S.D. It received support under the program "Investissements d'Avenir" launched by the French Government and implemented by ANR with the references ANR-10-LABX-54 MEMOLIFE and ANR-10-IDEX-0001-02 PSL Research. The sequencing effort was funded by France Génomique through involvement of the technical facilities of Genoscope (ANR-10-INBS-09-08). We acknowledge the ImagoSeine facility, member of the France BioImaging infrastructure supported by the ANR-10-INSB-04.

References

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

1175 doi:10.1371/journal.pgen.1008949
1176 9. Abello A, Régnier V, Arnaiz O, Le Bars R, Bétermier M, Bischerour J. Functional
1177 diversification of Paramecium Ku80 paralogs safeguards genome integrity during precise
1178 programmed DNA elimination. PLoS Genet. 2020;16: e1008723.
1179 doi:10.1371/journal.pgen.1008723
1180 10. Klobutcher LA, Herrick G. Consensus inverted terminal repeat sequence of
1181 Paramecium IESs: resemblance to termini of Tc1-related and Euplotes Tec transposons.
1182 Nucleic Acids Res. 1995;23: 2006–2013. doi:10.1093/nar/23.11.2006
1183 11. Klobutcher LA, Herrick G. Developmental genome reorganization in ciliated
1184 protozoa: the transposon link. Prog Nucleic Acid Res Mol Biol. 1997;56: 1–62.
1185 doi:10.1016/s0079-6603(08)61001-6
1186 12. Singh DP, Saudemont B, Guglielmi G, Arnaiz O, Goût J-F, Prajer M, et al. Genome-
1187 defence small RNAs exapted for epigenetic mating-type inheritance. Nature. 2014;509: 447–
1188 452. doi:10.1038/nature13318
1189 13. Sawka-Gądek N, Potekhin A, Singh DP, Grevtseva I, Arnaiz O, Penel S, et al.
1190 Evolutionary plasticity of mating-type determination mechanisms in Paramecium aurelia
1191 sibling species. Genome Biology and Evolution. 2021;In press. doi:10.1093/gbe/evaa258
1192 14. McGrath CL, Gout J-F, Johri P, Doak TG, Lynch M. Differential retention and
1193 divergent resolution of duplicate genes following whole-genome duplication. Genome
1194 Research. 2014;24: 1665–1675. doi:10.1101/gr.173740.114
1195 15. McGrath CL, Gout J-F, Doak TG, Yanagi A, Lynch M. Insights into Three Whole-
1196 Genome Duplications Gleaned from the *Paramecium caudatum* Genome Sequence. Genetics.
1197 2014;197: 1417–1428. doi:10.1534/genetics.114.163287
1198 16. Gout J-F, Johri P, Arnaiz O, Doak TG, Bhullar S, Couloux A, et al. Universal trends
1199 of post-duplication evolution revealed by the genomes of 13 Paramecium species sharing an
1200 ancestral whole-genome duplication. bioRxiv. 2019; 573576. doi:10.1101/573576
1201 17. Sonneborn TM. Paramecium aurelia. In: King RC, editor. Handbook of Genetics:
1202 Plants, Plant Viruses, and Protists. Boston, MA: Springer US; 1974. pp. 469–594.
1203 doi:10.1007/978-1-4684-2994-7_20
1204 18. Le Mouël A, Butler A, Caron F, Meyer E. Developmentally regulated chromosome
1205 fragmentation linked to imprecise elimination of repeated sequences in paramecia. Eukaryotic
1206 Cell. 2003;2: 1076–1090. doi:10.1128/ec.2.5.1076-1090.2003
1207 19. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and de novo assembly of
1208 the giant panda genome. Nature. 2010;463: 311–317. doi:10.1038/nature08696
1209 20. Duharcourt S, Sperling L. The Challenges of Genome-Wide Studies in a Unicellular
1210 Eukaryote With Two Nuclear Genomes. Meth Enzymol. 2018;612: 101–126.
1211 doi:10.1016/bs.mie.2018.08.012
1212 21. Johri P, Krenek S, Marinov GK, Doak TG, Berendonk TU, Lynch M. Population
1213 Genomics of Paramecium Species. Molecular Biology and Evolution. 2017;34: 1194–1216.
1214 doi:10.1093/molbev/msx074
1215 22. Denby Wilkes C, Arnaiz O, Sperling L. ParTIES: a toolbox for Paramecium
1216 interspersed DNA elimination studies. Bioinformatics. 2016;32: 599–601.
1217 doi:10.1093/bioinformatics/btv691
1218 23. Mayer KM, Mikami K, Forney JD. A mutation in Paramecium tetraurelia reveals
1219 functional and structural features of developmentally excised DNA elements. Genetics.
1220 1998;148: 139–149.
1221 24. Duharcourt S, Keller AM, Meyer E. Homology-dependent maternal inhibition of
1222 developmental excision of internal eliminated sequences in Paramecium tetraurelia. Mol Cell
1223 Biol. 1998;18: 7075–7085. doi:10.1128/mcb.18.12.7075
1224 25. Przyboś E, Tarcz S, Rautian M, Sawka N. Delimiting Species Boundaries within a

1225 Paraphyletic Species Complex: Insights from Morphological, Genetic, and Molecular Data on
1226 Paramecium sonneborni (Paramecium aurelia species complex, Ciliophora, Protozoa). Protist.
1227 2015;166: 438–456. doi:10.1016/j.protis.2015.07.001
1228 26. Boussau B, Szollosi GJ, Duret L, Gouy M, Tannier E, Daubin V. Genome-scale
1229 coestimation of species and gene trees. Genome Research. 2013;23: 323–330.
1230 doi:10.1101/gr.141978.112
1231 27. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo
1232 generator. Genome Res. 2004;14: 1188–1190. doi:10.1101/gr.849004
1233 28. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New
1234 algorithms and methods to estimate maximum-likelihood phylogenies: assessing the
1235 performance of PhyML 3.0. Syst Biol. 2010;59: 307–321. doi:10.1093/sysbio/syq010
1236 29. Duret L, Cohen J, Jubin C, Dessen P, Gout J-F, Mousset S, et al. Analysis of sequence
1237 variability in the macronuclear DNA of Paramecium tetraurelia: A somatic view of the
1238 germline. Genome Research. 2008;18: 585–596. doi:10.1101/gr.074534.107
1239 30. Lhuillier-Akakpo M, Frapporti A, Denby Wilkes C, Matelot M, Vervoort M, Sperling
1240 L, et al. Local Effect of Enhancer of Zeste-Like Reveals Cooperation of Epigenetic and cis-
1241 Acting Determinants for Zygotic Genome Rearrangements. PLoS Genetics. 2014;10:
1242 e1004665. doi:10.1371/journal.pgen.1004665
1243 31. Maliszewska-Olejniczak K, Gruchota J, Gromadka R, Denby Wilkes C, Arnaiz O,
1244 Mathy N, et al. TFIIS-Dependent Non-coding Transcription Regulates Developmental
1245 Genome Rearrangements. PLoS Genet. 2015;11: e1005383.
1246 doi:10.1371/journal.pgen.1005383
1247 32. Gruchota J, Denby Wilkes C, Arnaiz O, Sperling L, Nowak JK. A meiosis-specific
1248 Spt5 homolog involved in non-coding transcription. Nucleic Acids Res. 2017;45: 4722–4732.
1249 doi:10.1093/nar/gkw1318
1250 33. Sandoval PY, Swart EC, Arambasic M, Nowacki M. Functional diversification of
1251 Dicer-like proteins and small RNAs required for genome sculpting. Dev Cell. 2014;28: 174–
1252 188. doi:10.1016/j.devcel.2013.12.010
1253 34. Frapporti A, Miró Pina C, Arnaiz O, Holoch D, Kawaguchi T, Humbert A, et al. The
1254 Polycomb protein Ezl1 mediates H3K9 and H3K27 methylation to repress transposable
1255 elements in Paramecium. Nature Communications. 2019;10: 2710. doi:10.1038/s41467-019-
1256 10648-5
1257 35. Lepère G, Bétermier M, Meyer E, Duharcourt S. Maternal noncoding transcripts
1258 antagonize the targeting of DNA elimination by scanRNAs in Paramecium tetraurelia. Genes
1259 Dev. 2008;22: 1501–1512. doi:10.1101/gad.473008
1260 36. Arnaiz O, Van Dijk E, Bétermier M, Lhuillier-Akakpo M, de Vanssay A, Duharcourt
1261 S, et al. Improved methods and resources for paramecium genomics: transcription units, gene
1262 annotation and gene expression. BMC Genomics. 2017;18: 483. doi:10.1186/s12864-017-
1263 3887-z
1264 37. Blumenstiel JP. Birth, School, Work, Death, and Resurrection: The Life Stages and
1265 Dynamics of Transposable Element Proliferation. Genes. 2019;10: 336.
1266 doi:10.3390/genes10050336
1267 38. Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. Evidence for DNA Loss as
1268 a Determinant of Genome Size. Science. 2000;287: 1060–1062.
1269 doi:10.1126/science.287.5455.1060
1270 39. Mitra R, Fain-Thornton J, Craig NL. piggyBac can bypass DNA synthesis during cut
1271 and paste transposition. EMBO J. 2008;27: 1097–1109. doi:10.1038/emboj.2008.41
1272 40. Cavalier-Smith T. Intron phylogeny: a new hypothesis. Trends in Genetics. 1991;7:
1273 145–148. doi:10.1016/0168-9525(91)90377-3
1274 41. Wahl MC, Will CL, Lührmann R. The Spliceosome: Design Principles of a Dynamic

RNP Machine. Cell. 2009;136: 701–718. doi:10.1016/j.cell.2009.02.009

42. Doolittle WF. Is junk DNA bunk? A critique of ENCODE. Proceedings of the National Academy of Sciences. 2013;110: 5294–5300. doi:10.1073/pnas.1221376110

43. Huff JT, Zilberman D, Roy SW. Mechanism for DNA transposons to generate introns on genomic scales. Nature. 2016;538: 533–536. doi:10.1038/nature20110

44. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature. 2010;463: 457–463. doi:10.1038/nature08909

45. McGlincy NJ, Smith CWJ. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? Trends in Biochemical Sciences. 2008;33: 385–393. doi:10.1016/j.tibs.2008.06.001

46. Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature. 2007;446: 926–929. doi:10.1038/nature05676

47. Lareau LF, Brenner SE. Regulation of Splicing Factors by Alternative Splicing and NMD Is Conserved between Kingdoms Yet Evolutionarily Flexible. Molecular Biology and Evolution. 2015;32: 1072–1079. doi:10.1093/molbev/msv002

48. Pickrell JK, Pai AA, Gilad Y, Pritchard JK. Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. PLoS Genetics. 2010;6: e1001236. doi:10.1371/journal.pgen.1001236

49. Saudemont B, Popa A, Parmley JL, Rocher V, Blugeon C, Necsulea A, et al. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. Genome Biology. 2017;18: 208. doi:10.1186/s13059-017-1344-6

50. Tress ML, Abascal F, Valencia A. Most Alternative Isoforms Are Not Functionally Important. Trends in Biochemical Sciences. 2017;42: 408–410. doi:10.1016/j.tibs.2017.04.002

51. Lynch M. The Origins of Eukaryotic Gene Structure. Molecular Biology and Evolution. 2006;23: 450–468. doi:10.1093/molbev/msj050

52. Lynch M. Intron evolution as a population-genetic process. Proceedings of the National Academy of Sciences. 2002;99: 6118–6123. doi:10.1073/pnas.092595699

53. Hamilton EP, Kapusta A, Huvos PE, Bidwell SL, Zafar N, Tang H, et al. Structure of the germline genome of Tetrahymena thermophila and relationship to the massively rearranged somatic genome. eLife. 2016;5. doi:10.7554/eLife.19090

54. Cheng C-Y, Young JM, Lin C-YG, Chao J-L, Malik HS, Yao M-C. The piggyBac transposon-derived genes TPB1 and TPB6 mediate essential transposon-like excision during the developmental rearrangement of key genes in Tetrahymena thermophila. Genes Dev. 2016;30: 2724–2736. doi:10.1101/gad.290460.116

55. Feng L, Wang G, Hamilton EP, Xiong J, Yan G, Chen K, et al. A germline-limited piggyBac transposase gene is required for precise excision in Tetrahymena genome rearrangement. Nucleic Acids Research. 2017;45: 9481–9502. doi:10.1093/nar/gkx652

56. Cheng C-Y, Vogt A, Mochizuki K, Yao M-C. A Domesticated piggyBac Transposase Plays Key Roles in Heterochromatin Dynamics and DNA Cleavage during Programmed DNA Deletion in Tetrahymena thermophila. MBoC. 2010;21: 1753–1762. doi:10.1091/mbc.e09-12-1079

57. Vogt A, Mochizuki K. A domesticated PiggyBac transposase interacts with heterochromatin and catalyzes reproducible DNA elimination in Tetrahymena. PLoS Genet. 2013;9: e1004032. doi:10.1371/journal.pgen.1004032

58. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. Science. 1980;284: 601–603.

59. Orgel LE, Crick FHC. Selfish DNA: the ultimate parasite. Science. 1980;284: 604–607.

60. Werren JH. Selfish genetic elements, genetic conflict, and evolutionary innovation. Proceedings of the National Academy of Sciences. 2011;108: 10863–10870. doi:10.1073/pnas.1102343108
61. Krenek S, Berendonk TU, Petzoldt T. Thermal performance curves of Paramecium caudatum: a model selection approach. Eur J Protistol. 2011;47: 124–137. doi:10.1016/j.ejop.2010.12.001
62. Lhuillier-Akakpo M, Guérin F, Frapporti A, Duharcourt S. DNA deletion as a mechanism for developmentally programmed centromere loss. Nucleic Acids Res. 2016;44: 1553–1565. doi:10.1093/nar/gkv1110
63. Cummings DJ. Isolation and partial characterization of macro- and micronuclei from Paramecium aurella. J Cell Biol. 1972;53: 105–115. doi:10.1083/jcb.53.1.105
64. Freiburg M. Isolation and characterization of macronuclei of Paramecium caudatum infected with the macronucleus-specific bacterium Holospora obtusa. J Cell Sci. 1985;73: 389–398.
65. Galbraith DW, Harkins KR, Maddox JM, Ayres NM, Sharma DP, Firoozabady E. Rapid Flow Cytometric Analysis of the Cell Cycle in Intact Plant Tissues. Science. 1983;220: 1049–1051. doi:10.1126/science.220.4601.1049
66. Bourge M, Brown SC, Siljak-Yakovlev S. Flow cytometry as tool in plant sciences, with emphasis on genome size and ploidy level assessment. Genetics & Applications. 2018;2: 1–12. doi:10.31383/ga.vol2iss2pp1-12
67. Dolezel J, Bartos J, Voglmayr H, Greilhuber J. Nuclear DNA content and genome size of trout and human. Cytometry A. 2003;51: 127–128. doi:10.1002/cyto.a.10013
68. Marie D, Brown SC. A cytometric exercise in plant DNA histograms, with 2C values for 70 species. Biol Cell. 1993;78: 41–51. doi:10.1016/0248-4900(93)90113-s
69. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27: 578–579. doi:10.1093/bioinformatics/btq683
70. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1: 18. doi:10.1186/2047-217X-1-18
71. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15: R46. doi:10.1186/gb-2014-15-3-r46
72. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359. doi:10.1038/nmeth.1923
73. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19: 1639–1645. doi:10.1101/gr.092759.109
74. Foissac S, Gouzy J, Rombauts S, Mathe C, Amselem J, Sterck L, et al. Genome Annotation in Plants and Fungi: EuGene as a Model Platform. Current Bioinformatics. 2008;3: 87–97.
75. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27: 764–770. doi:10.1093/bioinformatics/btr011
76. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421. doi:10.1186/1471-2105-10-421
77. Miele V, Penel S, Duret L. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics. 2011;12: 116. doi:10.1186/1471-2105-12-116
78. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30: 772–780. doi:10.1093/molbev/mst010

79. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32: 268–274. doi:10.1093/molbev/msu300

80. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, et al. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature. 2006;444: 171–178. doi:10.1038/nature05230

81. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007;56: 564–577. doi:10.1080/10635150701472164

82. Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, et al. RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. Syst Biol. 2016;65: 726–736. doi:10.1093/sysbio/syw021

83. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. Bioinformatics. 2013;29: 2487–2489. doi:10.1093/bioinformatics/btt403

84. Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). Genome Biol Evol. 2015;7: 1192–1205. doi:10.1093/gbe/evv050

Supporting information

S1 Fig. Multi-gate flow cytometry strategy for sorting the MICs.
GFP, DAPI-positive MICs from *P. sonneborni* vegetative cells transformed with the *P. tetraurelia CENH3a-GFP* transgene [62] were sorted based on size, granularity, DAPI staining and GFP signal (see Materials and Methods). P4 and P8 were sorted separately. Based on quality control by flow imaging (Imagestream) indicating 97% purity, the two samples P4 and P8, which represent 1.91% of total events, were combined for DNA extraction and sequencing. Two populations are visualized and likely correspond to 2n and 4n MICs.

S2 Fig. Example of a MAC-variable region
Circular representation of one scaffold of ~730kb. The tracks from the exterior to the interior of the circle: G+C content of 100nt sliding windows (black), MAC DNA-seq depth (purple), the density in predicted non-coding genes (orange), RNA-seq depth (red) and the density of detected telomerisation sites (green). The external blue arc shows the region identified as being MAC-variable. These regions were determined by an automatic pipeline (see Materials and Methods), then adjusted by eye for each scaffold.

S3 Fig. Comparison of cytometry and k-mer MIC genome size estimates.
Flow cytometry estimates of DNA content of micronuclei and k-mer counting estimates of genome size are described in Materials and Methods. In order to show all of the data, both axes of the graph are log-transformed. Simple linear regression was carried out on the untransformed data with R. The linear model that fits the data is presented as a dashed blue line; $R^2 = 0.99$, p-value = 1.3×10^{-09}.

S4 Fig. Repeat content of *P. caudatum* MIC genome.
(A) Abundance of repeat families identified by DNAPipeTE in *P. caudatum* strain My43c3d. The repeat content of the *P. caudatum* MIC genome was analyzed with DNAPipeTE [84], using a sample of 3,500,000 sequence reads (corresponding to a read depth of ~0.5X). DNAPipeTE identified 67 repeat families that collectively constitute 83% of the MIC genome. Among them, there are two major satellite repeats Sat1 and Sat2, which represent respectively 42% and 29% of the MIC genome.
(B) Sequences of the two major satellite repeats Sat1 and Sat2 in *P. caudataum* My43c3d (332 bp and 449 bp long). These two satellite repeats share homology over a ~200 bp-long region. Primer sequences used for specific PCR amplification of each repeat are indicated in bold.
(C) Detection of Sat1 and Sat2 *in P. caudatum* strains. Whole cell genomic DNA was used to perform duplex PCR with a set of primers located within each repeat (Sat1 or Sat2, in bold panel B) and another set of primers within the 18S ribosomal DNA as a loading control. The expected size of the 18SrDNA PCR product was 301 bp using primers 18S_F953: AGACGATCAGATACCGTCGTAG and 18S_R1300: CACCAACTAAGAACGGCCATGC. L: 1-kb NEB ladder. Neg.: negative control (no DNA). Sat1 was amplified with primers comp2975_F1: TTGTGCTGTAGGGCTCAATAAT and comp2975_R1: CTCAAAATTCGACGCTGACAA at the expected size (198 bp) in *the P. caudatum* clade B strains tested (My43c3d; C033; C083; C131; C147). The repeat could not be amplified in *P. caudatum* DNA from clade A strains (C023; C065; C104; C119), from strain C026 or from strain Indo_1.6I.
Sat2 was amplified with primers comp5240_F1: TGCTGCTGATTTTGGATCTCG and comp5240_R1: CCGAGAACGGCCATTACAAG at the expected size (168 bp) in the *P. caudatum* clade B strains tested (My43c3d; C033; C083; C131; C147). The repeat could not

be amplified in *P. caudatum* DNA from clade A strains (C023; C065; C104; C119), from strain C026 or from strain Indo_1.6I.

S5 Fig. Intragenic IES density vs. gene expression level.
Expression levels (RPKM) were measured with RNAseq datasets from vegetative cells. For each species, expressed genes were classified into 10 bins of equal sample size according to their expression level, and we computed the IES density within each bin. Non-expressed genes (6.6% of the entire dataset) were excluded. (A) *Paramecium aurelia* species. (B) *P. caudatum*.

S6 Fig. Dating events of IES insertion/loss.
(A) To date events of IES loss or gain, it is first necessary to identify IESs that are homologous. For this, we aligned coding sequences (based on the protein alignment) and mapped the position of IESs: IESs located at the exact same position within a codon were assumed to be homologous (i.e. to result from a single ancestral insertion event). We then used the reconciled gene tree to map events in the species phylogeny, using a maximum likelihood approach (see methods). The example shown here corresponds to a gene family encoding a putative RNA 3'-terminal phosphate cyclase (PTET.51.1.P0920097, POCTA.138.1.P0960088, PBIA.V1_4.1.P01950012, PTRED.209.2.P71800001293600070, PPENT.87.1.P1090087, PPRIM.AZ9-3.1.P0020612, PSON.ATCC_30995.1.P0860097, PSEX.AZ8_4.1.P0910047, PCAU.43c3d.1.P00760109). The positions of IESs are indicated by red rectangles. (B) The presence of IESs (red bars) within each of these genes is indicated with regard to the species phylogeny. Six distinct IESs were identified in this gene family: IES2 is shared by all species and therefore predates the divergence between *P. caudatum* and the *aurelia* clade; IES4 most probably corresponds to a gain in the *P. sexaurelia* lineage; IES5 and IES6 predate the divergence of the *aurelia* clade and have been subsequently lost in the *P. tetraurelia/P. octaurelia* lineage; IES1 might correspond to a gain at the base of the *aurelia* clade or a loss in the *P. caudatum* lineage (and vice versa for IES3).

S7 Fig. Prevalence of weak IESs.
(A) Proportion of weak IESs (i.e. IESs with a retention frequency ≥10% in WT vegetative cells) among IESs located in different genomic compartments. (B) Proportion of weak IESs according to the age of IESs (for IESs located in coding regions): New = species-specific IES; Old = IES predating the divergence between *P. caudatum* and the *aurelia* lineage. The number of new IESs is indicated for each species. 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*

S8 Fig. Length distribution of IESs according to their age.
Comparison of the length distribution of IESs according to their age (for the subset of datable IESs located in coding regions). The age of an IES site is defined as in Fig. 3. Results for other species are shown in Fig. 4.

S9 Fig. Genomic distribution of IESs according to their length.
Green bars indicate the percentage of IESs located in each compartment of the MAC genome (introns, protein-coding regions, intergenic regions) for each species. Grey bars indicate the percentage of the MAC genome in each compartment. (A) Long IESs (>100 bp). (B) Short IESs (<35 bp).

S10 Fig. A highly conserved IES contributes to the regulation of gene expression.

The IES family FAM_9405 is present at the 5' end of a protein-coding gene of unknown function, expressed at a high level, specifically during autogamy. The IES overlaps the beginning of the first exon, including the 5'UTR and the first codons. Excision of the IES during MAC development leads to the loss of the initiation codon and of the promoter region, and thereby to the silencing of this gene in vegetative cells. (A) Gene structure and expression level during autogamy in *P. tetraurelia*. The IES is displayed in red. The position of the translation start site is indicated by a red arrow. (B) Alignment of homologous IESs across *aurelia* species. The N-terminal end of the encoded protein is shown below.

S11 Fig. The vast majority of IESs correspond to unique sequences.
For each species, all IESs were compared against each other with BLASTN (with an E-value threshold of 10^{-5}). The distribution of the number of BLAST hits per IES (excluding self-hits) is displayed for each species.

S12 Fig. Estimating the proportion of MIC and MAC DNA in the sample based on IES retention score.
The histograms on the left show the k-mer depth profiles. The peak at the origin can be attributed to sequencing errors (k-mers that occur only once or a few times). The position of the largest peak beyond the origin corresponds to k-mers present once in the genome and provides the sequencing depth. As *Paramecium aurelia* genomes have undergone whole genome duplications, there are a significant number of k-mers at 2X and even 4X the sequencing depth arising from genes (or regions of genes) present in 2 or 4 copies, clearly visible for *P. octaurelia* and *P. primaurelia*. The profile for *P. tetraurelia* however has a first peak (MIC sequences that occur once) at 31X followed by a larger peak that is not at the 2X position as it arises because of MAC DNA contamination. The column on the right shows histograms of IES retention scores. Only the *P. tetraurelia* sample is significantly contaminated by MAC DNA: the average IES retention score of 0.4 indicates 40% MIC and 60% MAC DNA in this sample.

S13 Fig. Example of floating IES.
Comparison of MIC and MAC sequences indicates the presence of an IES at this locus. However, because of the presence of a repeated motif at the boundaries of the IES (blue arrows), it is not possible to determine which of the two possible segments (IES-1 in black or IES-2 in red) is actually excised *in vivo*. Such IESs that cannot be unambiguously positioned are called 'floating IESs'. They represent 6.8% of the 400,254 IESs detected across all species. In the vast majority of cases (86%) the alternative locations of floating IESs differ by only two bp (as in the example shown here), and there are less than 1% of floating IESs for which the uncertainty in IES position exceeds 5 bp.

S14 Fig. Measuring the rate of IES gain or loss along the species phylogeny.
To illustrate our methodology, we show here an example of a gene family with 3 genes, two from *P. sonneborni* (pson1, pson2) and one from *P. sexaurelia* (psex1). Two IES loci are found in this family (A, B). The probability of presence of an IES (estimated by Bayesian ancestral state reconstruction - see methods) is indicated by shaded circles for each locus at each node of the gene phylogeny. We focus here on the branch of the species tree leading from the common ancestor of *P. sexaurelia* and *P. sonneborni* to the leaf node of *P. sonneborni* (the red branch in the species tree, shown in insert). The length of this branch (*b*), is taken as a proxy for time. Because of a duplication event, this branch of the species tree corresponds to two paths in the gene tree (k=2). To estimate the IES gain rate, we calculate for each path the sum of increase in the probability of presence of an IES, for all IES loci (p^+). Along the first path (from the root

to pson1), we have p^+A1=0.5 and p^+B1=0. Along the second path (from the root to pson2), we have p^+A2=0.5 and p^+B2=0. The average gain rate along all paths, per unit 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 the number of well aligned sites in the gene family alignment (i.e. the number of sites where the presence of homologous IESs can be assessed). Similarly, to estimate the IES loss rate, we calculate for each path the sum of decrease in the probability of presence of an IES, for all IES loci (p^-). Along the first path (from the root to pson1), we have p^-A1=0 and p^-B1=0.4. Along the second path (from the root to pson2), we have p^-A2=0 and p^-B2=0.4. The average gain rate along all paths, per unit of time and per bp, is thus given by $L=(p^-A1 + p^-B1 + p^-A2 + p^-B2)/(k \times b \times I)$, where I is the number of IES loci in the gene family (here I=2).

S1 Table: MIC genome sequencing data.

S2 Table. MAC genome assemblies used in this study.
The assemblies of the four MAC genomes sequenced in the course of this project include both 'constitutive MAC' regions (i.e. regions that are always retained in the MAC) and 'MAC-variable regions' (i.e. regions that are mostly restricted to the MIC, but that are retained at low frequency in MAC nuclei). The size and content of these two types of regions are indicated.

S3 Table. Distribution of IESs in different genomic compartments.
Values in parentheses indicate the proportions expected under the hypothesis of uniform IES distribution along MAC-destined regions.

S4 Table. Highly conserved IESs that are transcribed during MAC development and/or associated to genes that are upregulated during MAC development. The transcription level of IESs is indicated for different stages during MAC development (S, T0 to T45) and in vegetative cells (V) [36]. Only *P. tetraurelia* IESs are shown in the table, because this is the only species for which developmental transcriptome data is available [36]. (*) The IES pte.MICA.16.324097 (FAM_4968) contains a complete protein-coding gene, which is expressed during development (HTH CenpB-type DNA-binding domain see Fig. 5). The gene in which this IES is inserted (PTET.51.1.P0160202) is not specifically expressed during development.

S5 Table. Sequencing data generated for this study.

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

S2 Data. Locations of MAC-variable regions and MAC assembly curation
This file provides the positions of MAC-variable regions identified in the MAC assemblies of *P. octaurelia, P. pentaurelia, P. primaurelia,* and *P. sonneborni* . In addition, it indicates the positions of putative assembly chimeras that have been identified in *P. octaurelia*, *P. primaurelia* and *P. sexaurelia*.

Figures

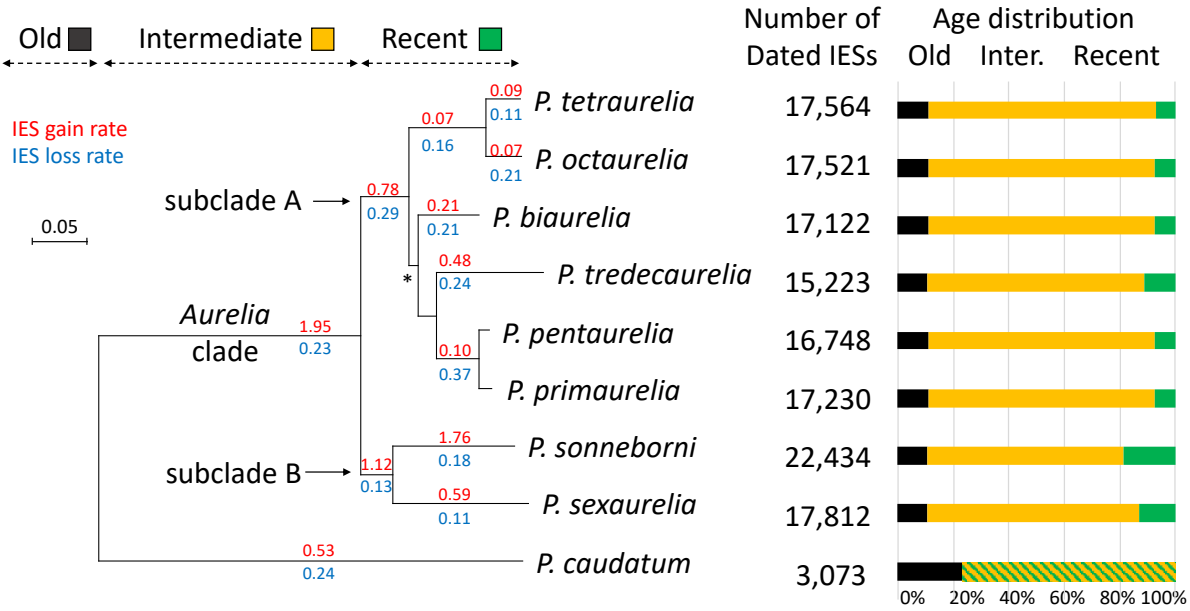


Fig 1. Dynamics of IES insertion/loss in *Paramecium*.
The species phylogeny was reconstructed from a concatenated alignment of 1,061 single-copy genes. All internal branches are supported by 100% bootstrap values (except branch *: bootstrap support = 83%). The age of IESs located within coding regions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

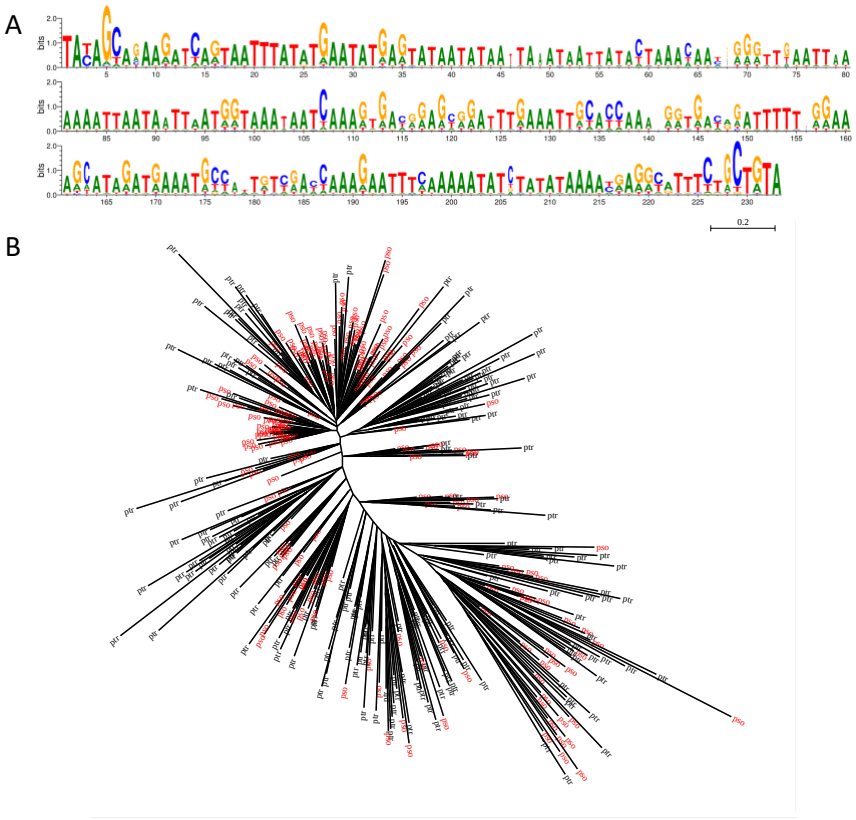


Fig 2. Phylogenetic analysis of the largest family of mobile IESs.
(A) Sequence logo [26], based on the alignment of the entire FAM_2183 family (N=4,153 mobile IESs). All copies present a high level of sequence similarity (average pairwise identity=72%) throughout their entire length (233 bp), not just at their ends. (B) 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 [27]). The tree topology is mainly star-like, which indicates that most copies derive from several bursts of insertions.

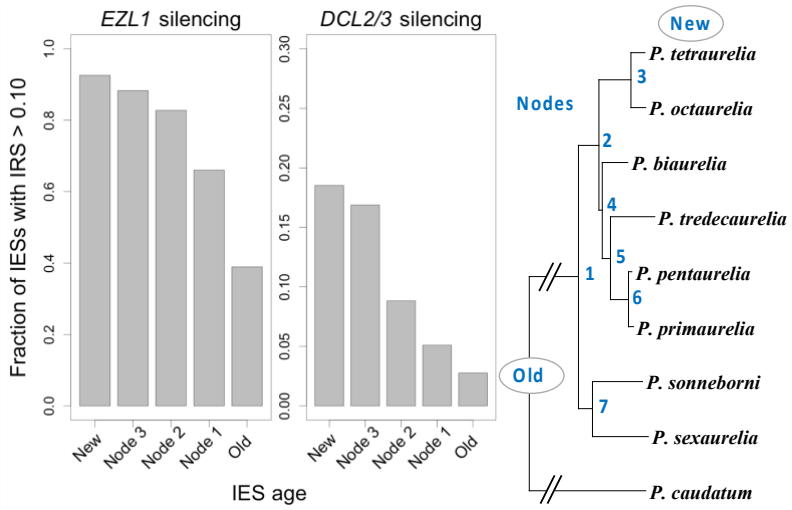


Fig 3. Older IESs are less dependent on Ezl1 and Dcl2/Dcl3 for their excision.
Barplots represent the fraction of *P. tetraurelia* IESs with a high retention score (IRS > 10%) after silencing of *EZL1* or *DCL2/DCL3*, according to their age. The age of an IES insertion is defined by the phylogenetic position of the last common ancestor (LCA) of species sharing an IES at the same site (New: *P. tetraurelia*-specific IES; Node *n*: the LCA corresponds to node number *n* in the species phylogeny; Old: the LCA predates the *P. aurelia*/*P. caudatum* divergence).

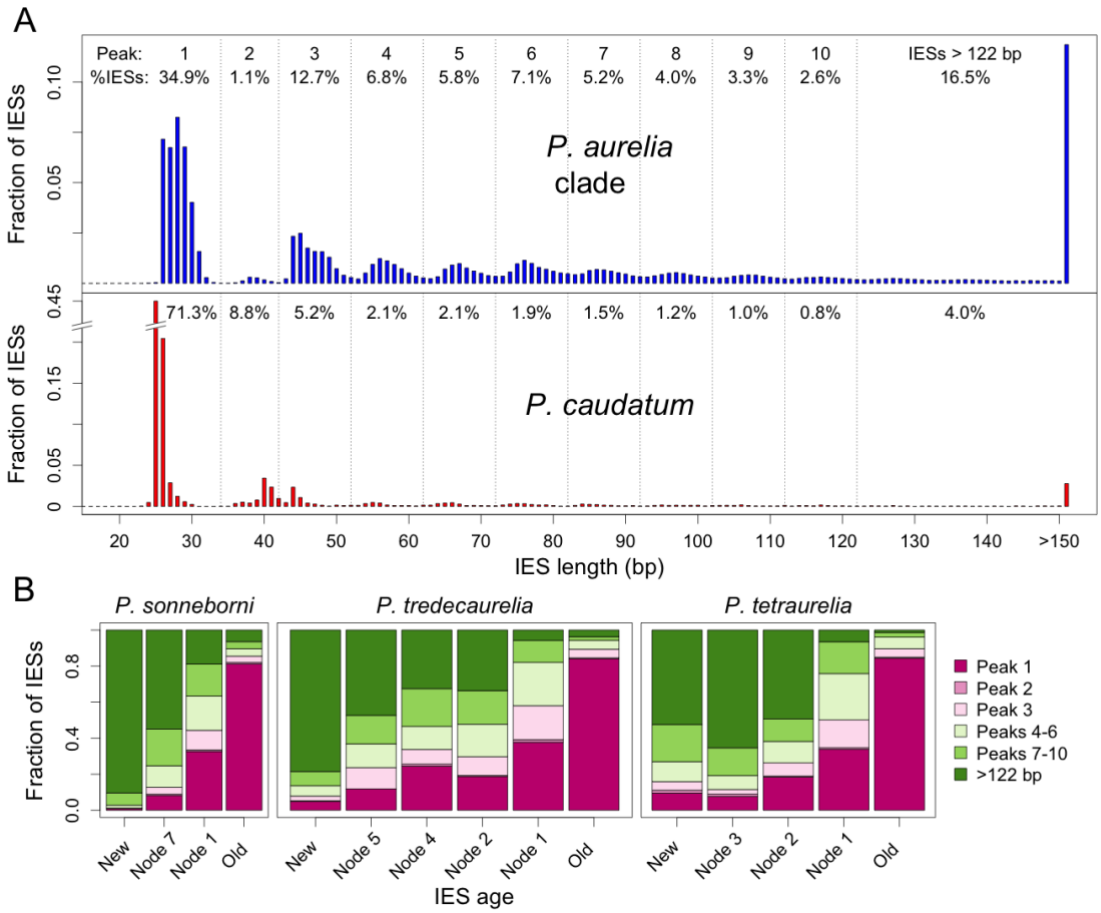


Fig 4. Length distribution of IESs according to their age.
(A) Comparison of the length distribution of IESs in *P. caudatum* (N=8,172 IESs) and in species from the *aurelia* clade (N=392,082 IESs). The fraction of IESs present within each peak of the distribution is indicated for the first 10 peaks. (B) Comparison of the length distribution of IESs according to their age (for the subset of datable IESs located in coding regions). The age of IES insertions is defined as in Fig 3. Results from other species are presented in S8 Fig.

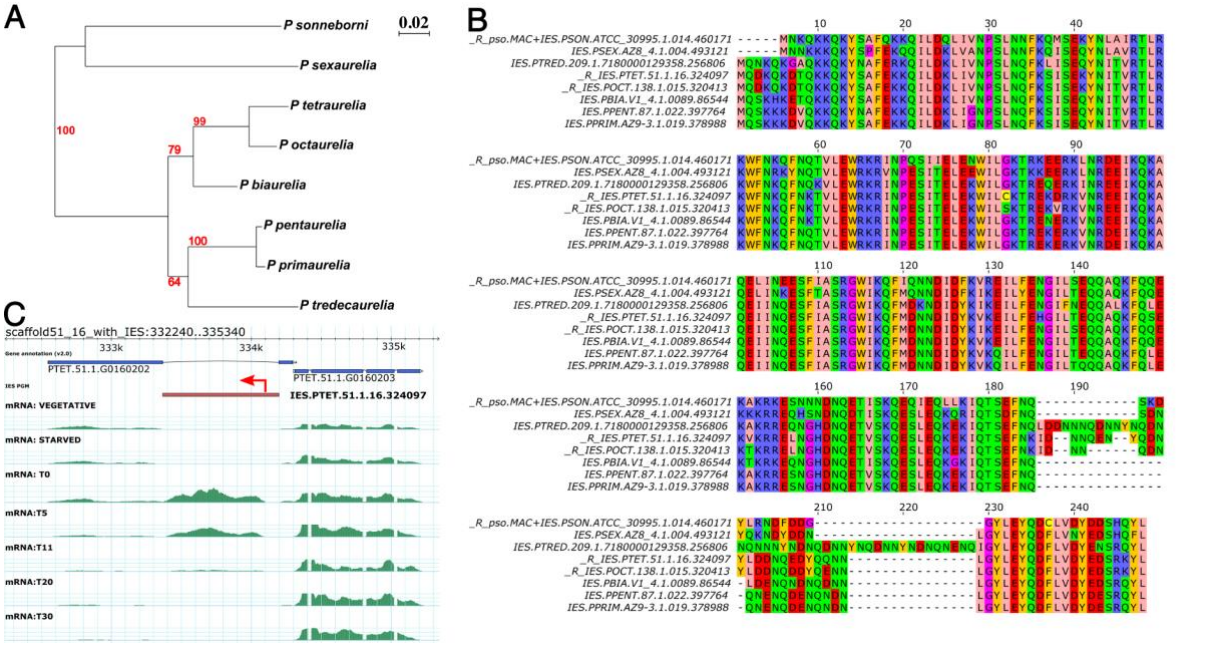
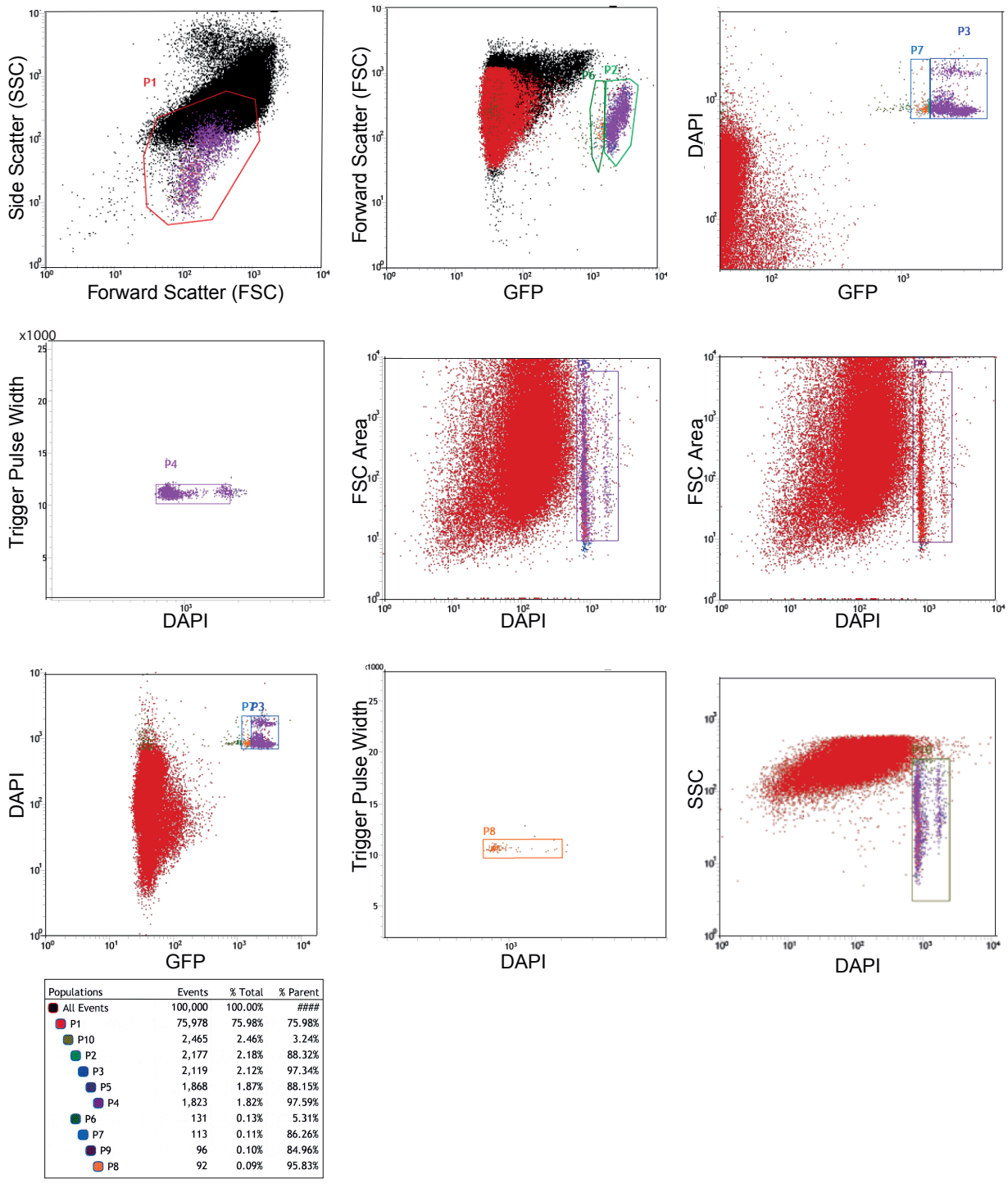


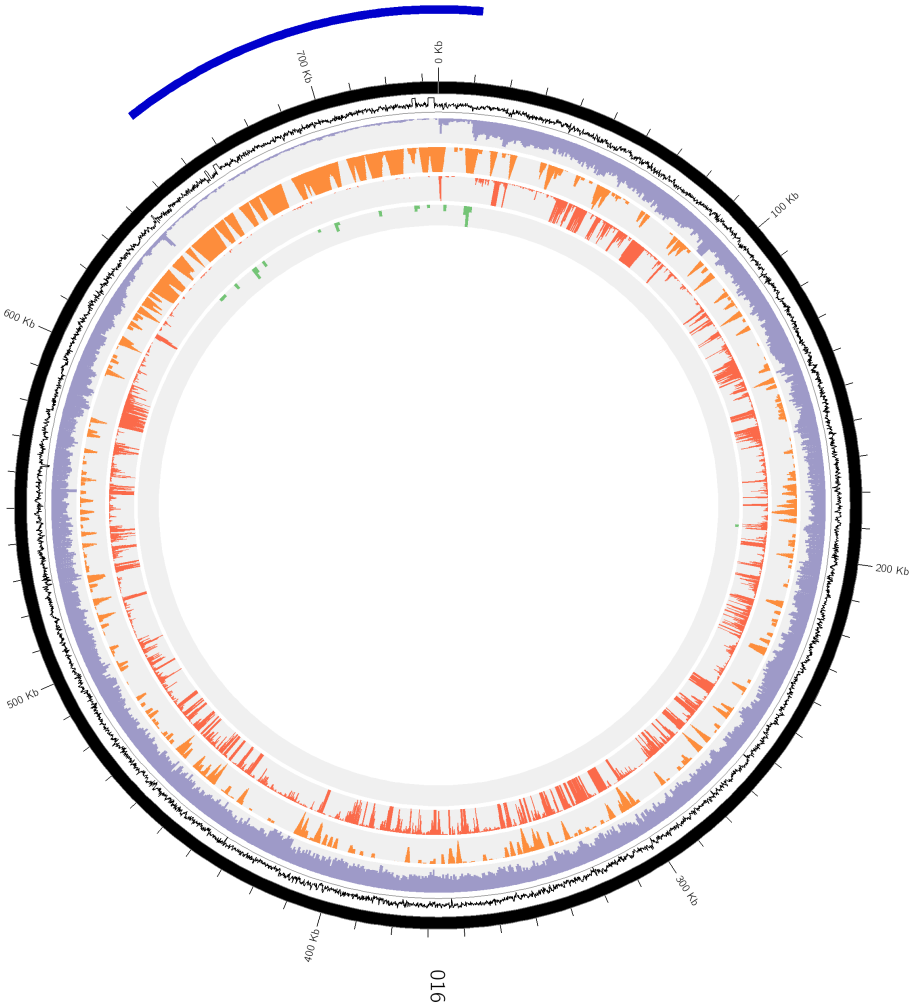
Fig 5. A highly conserved IES contains a gene encoding a putative DNA-binding protein. (A) Phylogenetic tree of the IES family FAM_4968. This IES is highly conserved in all species of the *aurelia* clade (≥75% nucleotide identity between the most distantly related species). BioNJ tree for 211 sites, Poisson model, 100 replicates. (B) Multiple alignment of the protein encoded by this IES. The coding region is subject to strong purifying selection (dN/dS=0.14). The encoded protein contains a helix-turn-helix DNA-binding domain (PF03221, IPR006600). (C) Gene annotation and expression level during autogamy of *P. tetraurelia*. The IES is located within a gene (on the opposite strand). The gene within the IES is expressed at high levels during the early stages of autogamy (T0 and T5).

Supplementary Figures



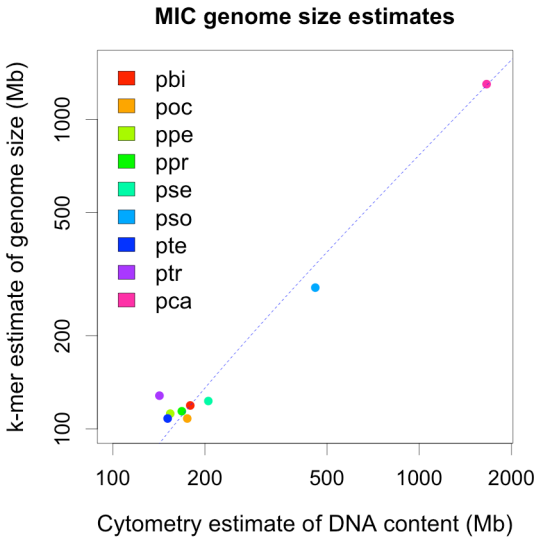
Populations	Events	% Total	% Parent
All Events	100,000	100.00%	####
P1	75,978	75.98%	75.98%
P10	2,465	2.46%	3.24%
P2	2,177	2.18%	88.32%
P3	2,119	2.12%	97.34%
P5	1,868	1.87%	88.15%
P4	1,823	1.82%	97.59%
P6	131	0.13%	5.31%
P7	113	0.11%	86.26%
P9	96	0.10%	84.96%
P8	92	0.09%	95.83%

S1 Fig. Multi-gate flow cytometry strategy for sorting the MICs.
GFP, DAPI-positive MICs from *P. sonneborni* vegetative cells transformed with the *P. tetraurelia* CENH3a-GFP transgene [60] were sorted based on size, granularity, DAPI staining and GFP signal (see Materials and Methods). P4 and P8 were sorted separately. Based on quality control by flow imaging (Imagestream) indicating 97% purity, the two samples P4 and P8, which represent 1.91% of total events, were combined for DNA extraction and sequencing. Two populations are visualized and likely correspond to 2n and 4n MICs.

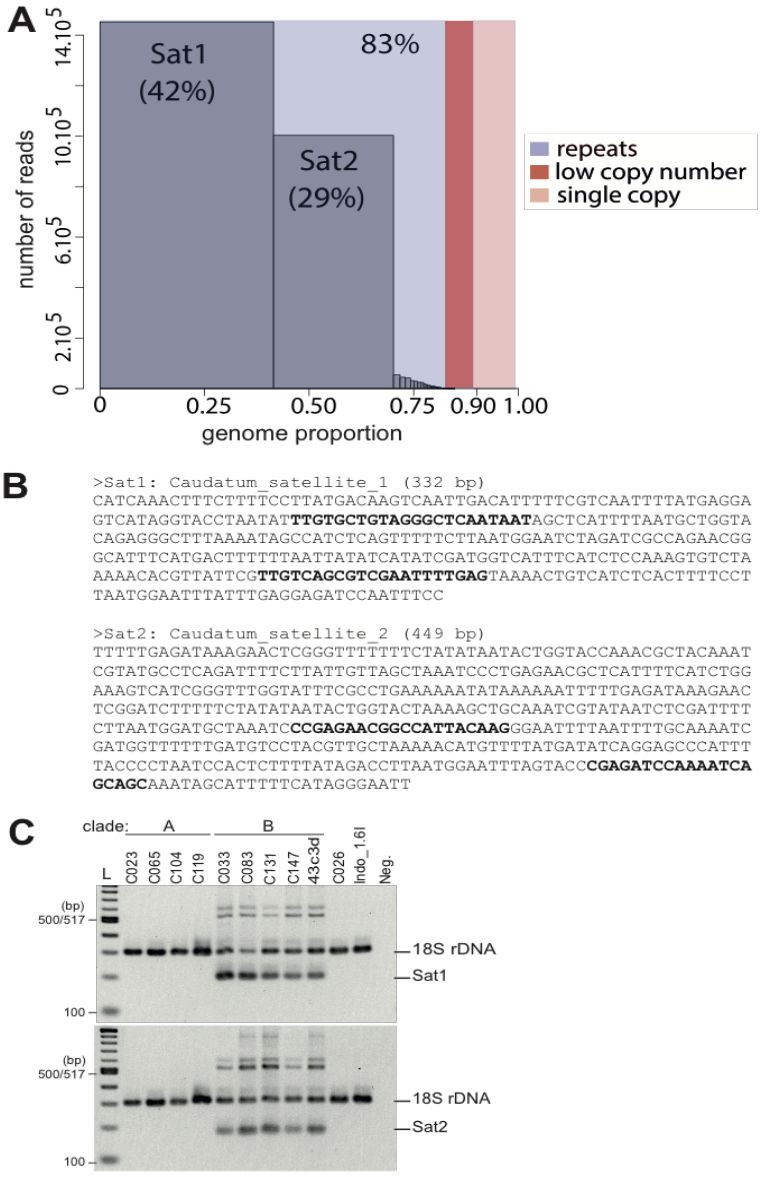


S2 Fig. Example of a MAC-variable region

Circular representation of one scaffold of ~730kb. The tracks from the exterior to the interior of the circle: G+C content of 100nt sliding windows (black), MAC DNA-seq depth (purple), the density in predicted non-coding genes (orange), RNA-seq depth (red) and the density of detected telomerisation sites (green). The external blue arc shows the region identified as being MAC-variable. These regions were determined by an automatic pipeline (see Materials and Methods) then adjusted by eye for each scaffold.



S3 Fig. Comparison of cytometry and k-mer MIC genome size estimates.
Flow cytometry estimates of DNA content of micronuclei and k-mer counting estimates of genome size are described in Materials and Methods. In order to show all of the data, both axes of the graph are log-transformed. Simple linear regression was carried out on the untransformed data with R. The linear model that fits the data is presented as a dashed blue line; $R^2 = 0.99$, p-value = 1.3×10^{-09}.



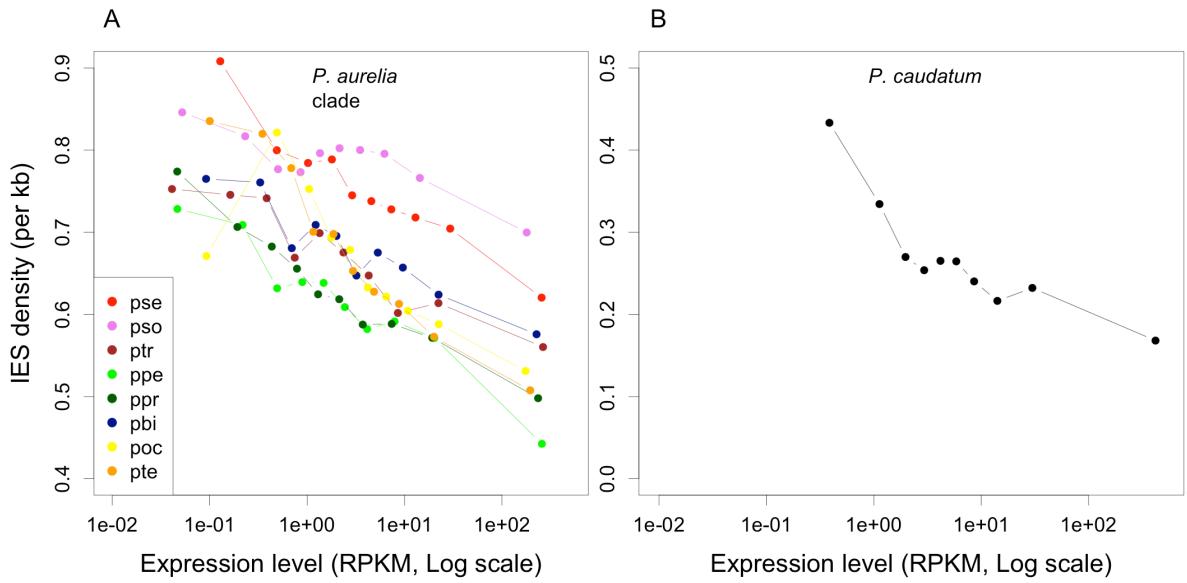
S4 Fig. Repeat content of *P. caudatum* MIC genome.

(A) Abundance of repeat families identified by DNAPipeTE in *P. caudatum* strain My43c3d. The repeat content of *P. caudatum* MIC genome was analyzed with DNAPipeTE [82], using a sample of 3,500,000 sequence reads (corresponding to a read depth of ~0.5X). DNAPipeTE identified 67 repeat families that collectively constitute 83% of the MIC genome. Among them, there are two major satellite repeats Sat1 and Sat2, which represent respectively 42% and 29% of the MIC genome.

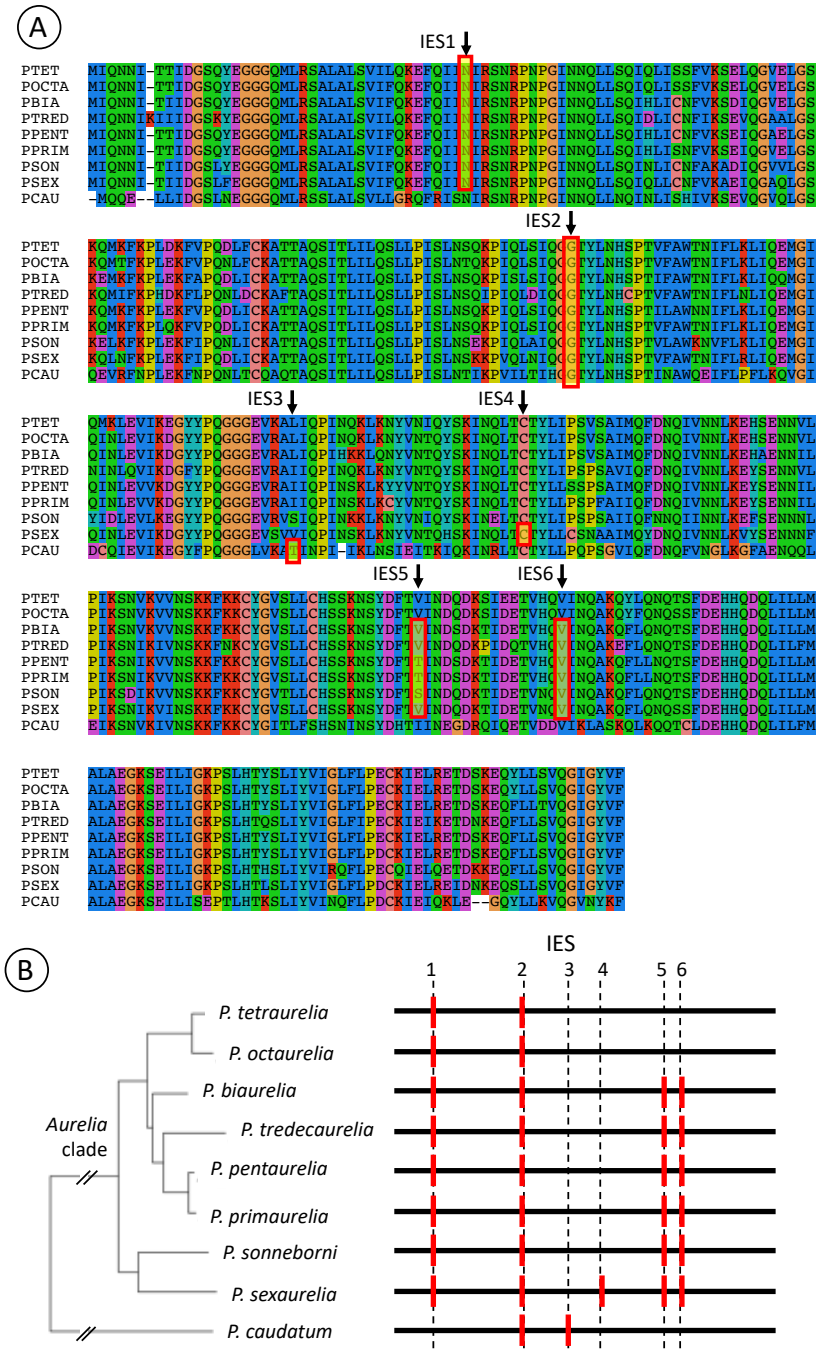
(B) Sequences of the two major satellite repeats Sat1 and Sat2 in *P. caudataum* My43c3d (332 bp and 449 bp long). These two satellite repeats share homology over a ~200 bp-long region. Primer sequences used for specific PCR amplification of each repeat are indicated in bold.

(C) Detection of Sat1 and Sat2 *in P. caudatum* strains. Whole cell genomic DNA was used to perform duplex PCR with a set of primers located within each repeat (Sat1 or Sat2, in bold panel B) and another set of primers within the 18S ribosomal DNA as a loading control. The expected size of the 18SrDNA PCR product was 301 bp using primers 18S_F953: AGACGATCAGATACCGTCGTAG and 18S_R1300: CACCAACTAAGAACGGCCATGC. L: 1-kb NEB ladder. Neg.: negative control (no DNA). Sat1 was amplified with primers comp2975_F1: TTGTGCTGTAGGGCTCAATAAT and comp2975_R1: CTCAAAATTCGACGCTGACAA at the expected size (198 bp) in *the P. caudatum* clade B strains tested (My43c3d; C033; C083; C131; C147). The repeat could not be amplified in *P. caudatum* DNA from clade A strains (C023; C065; C104; C119), from strain C026 or from strain Indo_1.6I.

Sat2 was amplified with primers comp5240_F1: TGCTGCTGATTTTGGATCTCG and comp5240_R1: CCGAGAACGGCCATTACAAG at the expected size (168 bp) in the *P. caudatum* clade B strains tested (My43c3d; C033; C083; C131; C147). The repeat could not be amplified in *P. caudatum* DNA from clade A strains (C023; C065; C104; C119), from strain C026 or from strain Indo_1.6I.



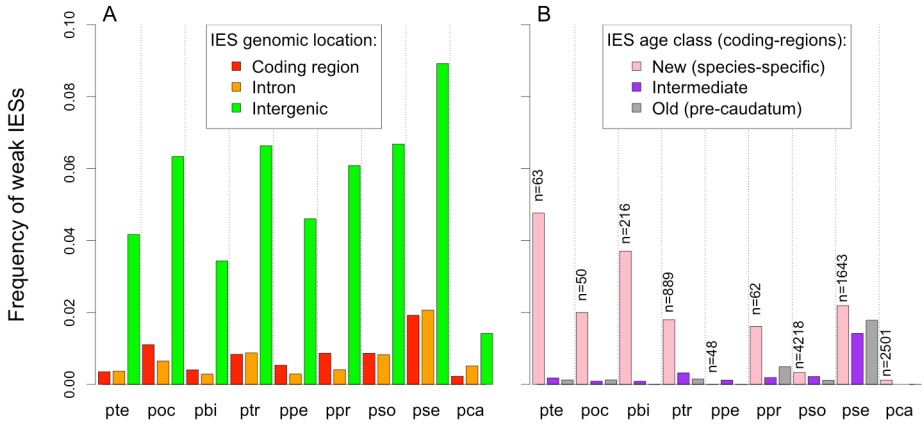
S5 Fig. Intragenic IES density vs. gene expression level.
Expression levels (RPKM) were measured with RNAseq datasets from vegetative cells. For each species, expressed genes were classified into 10 bins of equal sample size according to their expression level, and we computed the IES density within each bin. Non-expressed genes (6.6% of the entire dataset) were excluded. (A) *Paramecium aurelia* species. (B) *P. caudatum*.



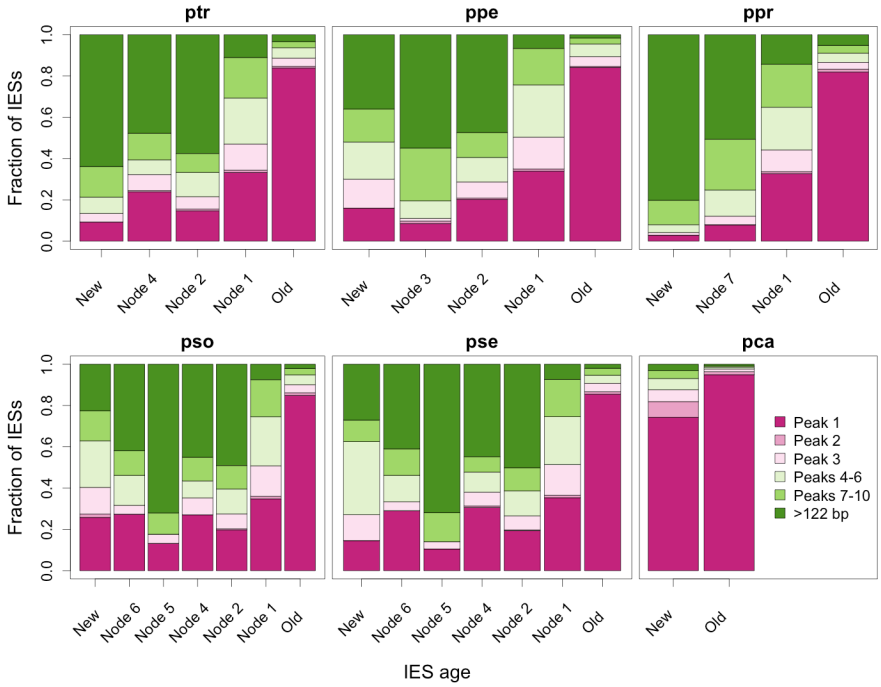
S6 Fig. Dating events of IES insertion/loss.

(**A**) To date events of IES loss or gain, it is first necessary to identify IESs that are homologous. For this, we aligned coding sequences (based on the protein alignment) and mapped the position of IESs: IESs located at the exact same position within a codon were assumed to be homologous (i.e. to result from a single ancestral insertion event). We then used the reconciled gene tree to map events in the species phylogeny, using a maximum likelihood approach (see methods). The example shown here corresponds to a gene family encoding a putative RNA 3'-terminal phosphate cyclase (PTET.51.1.P0920097, POCTA.138.1.P0960088, PBIA.V1_4.1.P01950012, PTRED.209.2.P71800001293600070, PPENT.87.1.P1090087, PPRIM.AZ9-3.1.P0020612, PSON.ATCC_30995.1.P0860097, PSEX.AZ8_4.1.P0910047, PCAU.43c3d.1.P00760109). The positions of IESs are indicated by red rectangles.

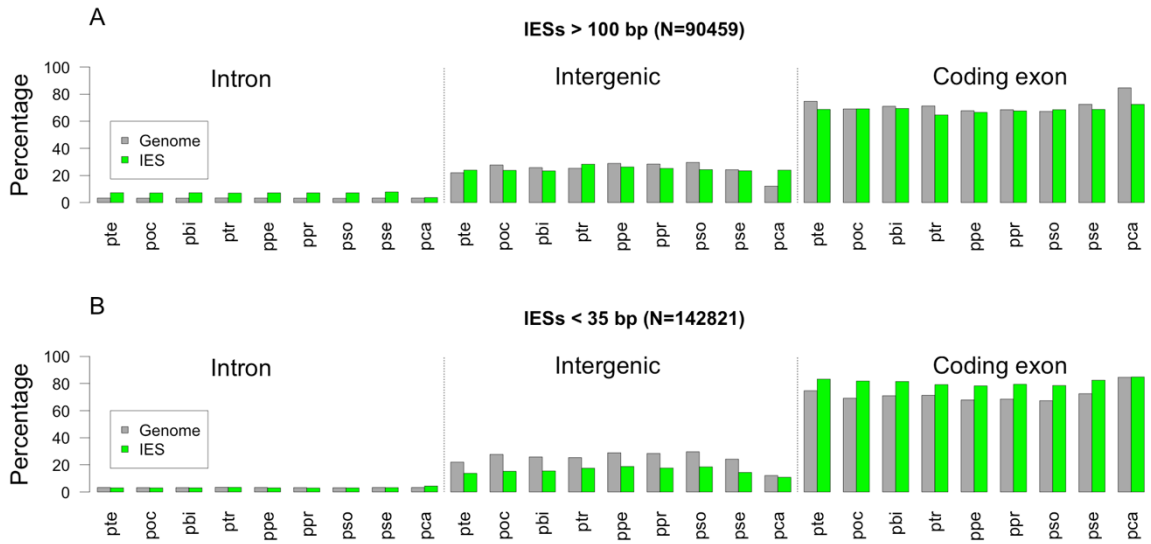
(**B**) The presence of IESs (red bars) within each of these genes is indicated with regard to the species phylogeny. Six distinct IESs were identified in this gene family: IES2 is shared by all species and therefore predates the divergence between *P. caudatum* and the *aurelia* clade; IES4 most probably corresponds to a gain in the *P. sexaurelia* lineage; IES5 and IES6 predate the divergence of the *aurelia* clade and have been subsequently lost in the *P. tetraurelia/P. octaurelia* lineage; IES1 might correspond to a gain at the base of the *aurelia* clade or a loss in the *P. caudatum* lineage (and vice versa for IES3).



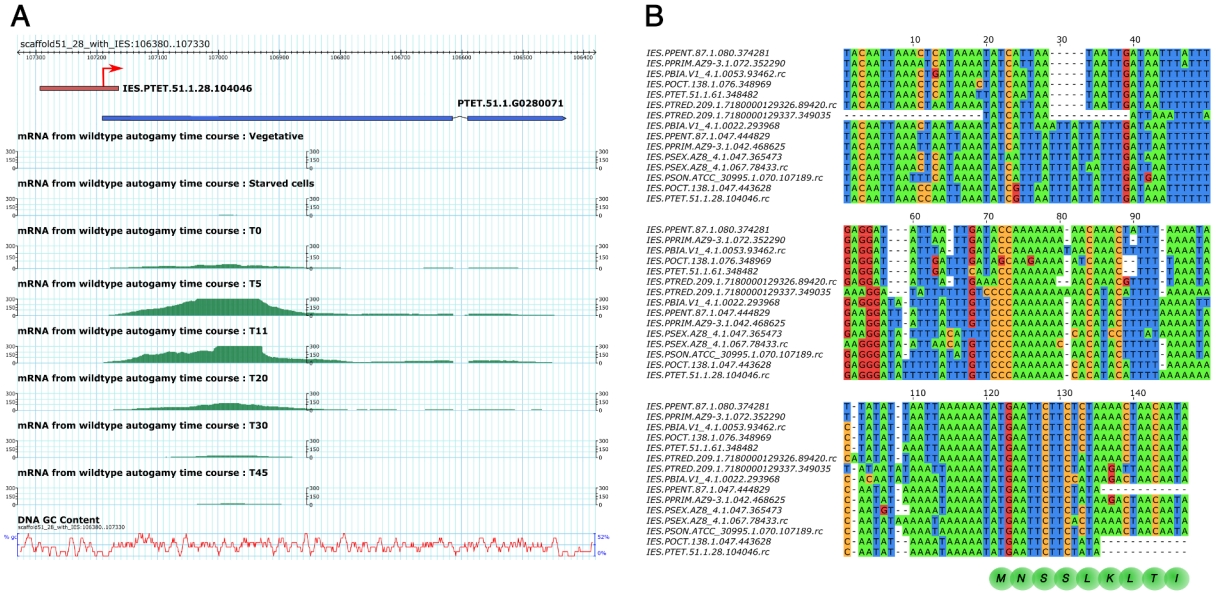
S7 Fig. Prevalence of weak IESs.
(**A**) Proportion of weak IESs (i.e. IESs with a retention frequency ≥10% in WT vegetative cells) among IESs located in different genomic compartments.
(**B**) Proportion of weak IESs according to the age of IESs (for IESs located in coding regions): New = species-specific IES; Old = IES predating the divergence between *P. caudatum* and the *aurelia* lineage. The number of new IESs is indicated for each species.
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*



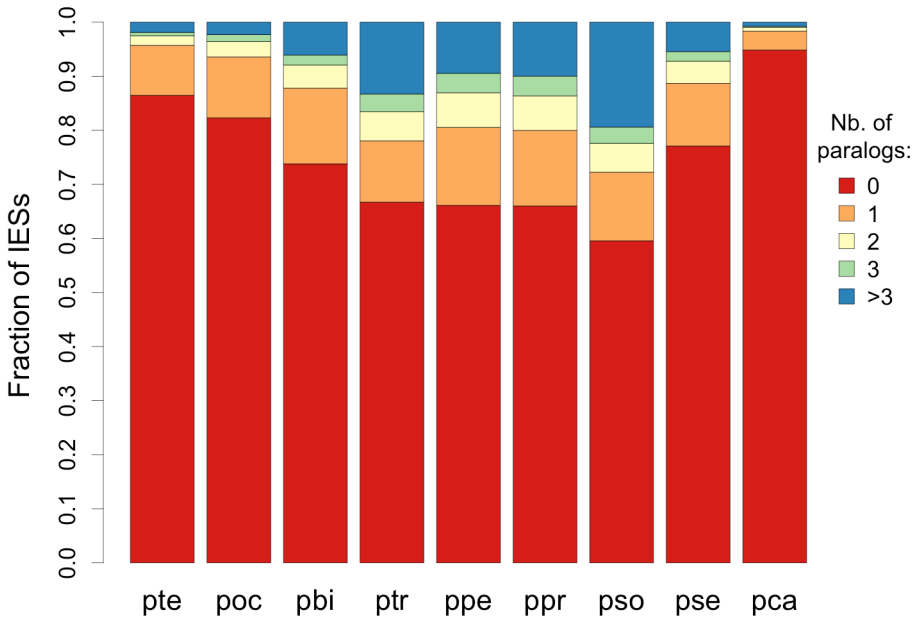
S8 Fig. Length distribution of IESs according to their age.
Comparison of the length distribution of IESs according to their age (for the subset of datable IESs located in coding regions). The age of an IES site is defined as in Fig. 3. Results for other species are shown in Fig. 4.



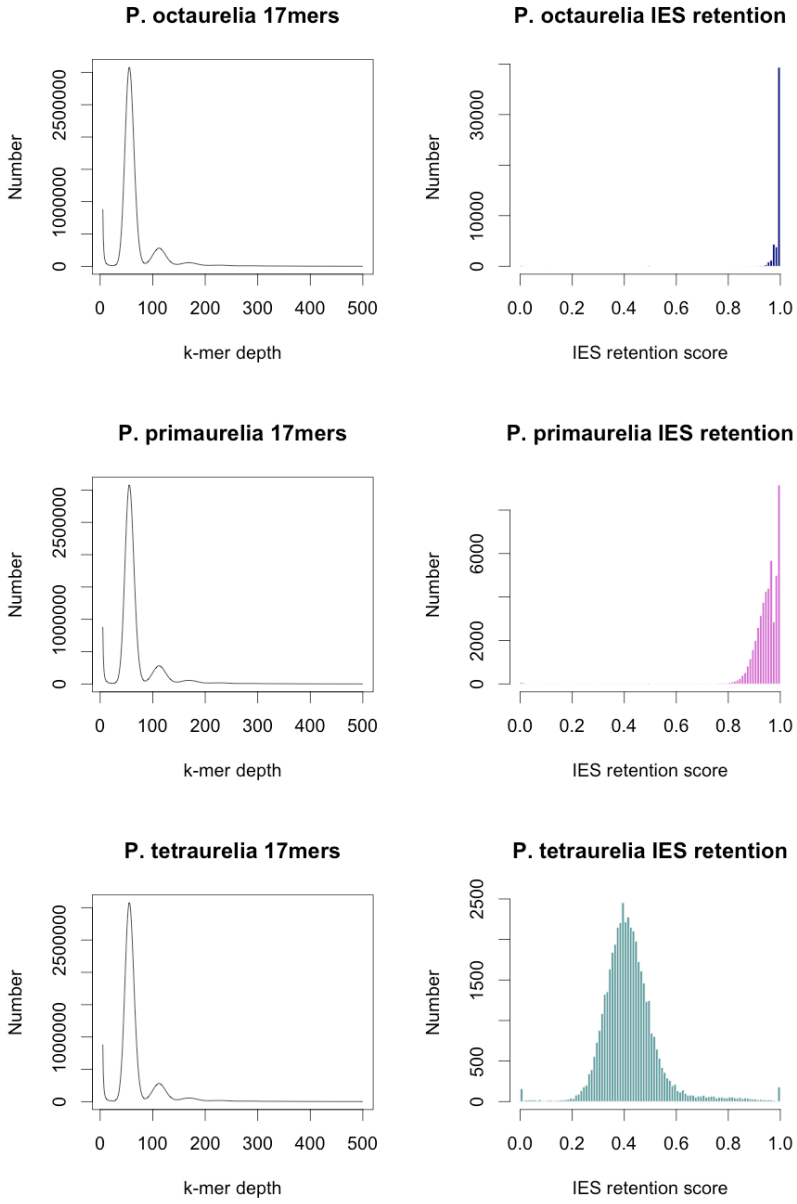
S9 Fig. Genomic distribution of IESs according to their length.
Green bars indicate the percentage of IESs located in each compartment of the MAC genome (introns, protein-coding regions, intergenic regions) for each species. Grey bars indicate the percentage of the MAC genome in each compartment. (A) Long IESs (>100 bp). (B) Short IESs (<35 bp).



S10 Fig. A highly conserved IES contributes to the regulation of gene expression.
The IES family FAM_9405 is present at the 5' end of a protein-coding gene of unknown function, expressed at a high level, specifically during autogamy. The IES overlaps the beginning of the first exon, including the 5'UTR and the first codons. Excision of the IES during MAC development leads to the loss of the initiation codon and of the promoter region, and thereby to the silencing of this gene in vegetative cells. (**A**) Gene structure and expression level during autogamy in *P. tetraurelia*. The IES is displayed in red. The position of the translation start site is indicated by a red arrow. (**B**) Alignment of homologous IESs across *aurelia* species. The N-terminal end of the encoded protein is shown below.

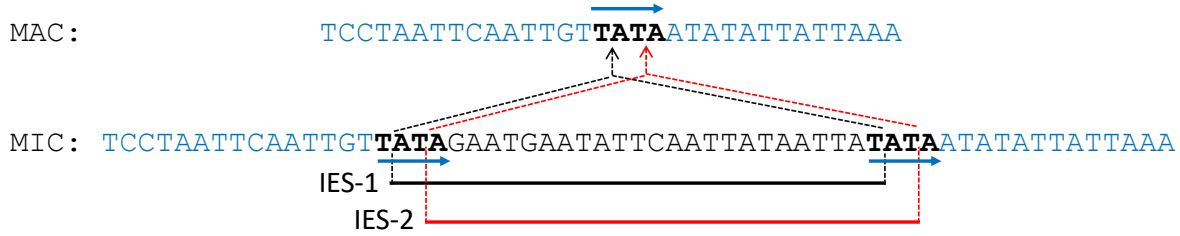


S11 Fig. The vast majority of IESs correspond to unique sequences.
For each species, all IESs were compared against each other with BLASTN (with an E-value threshold of 10^{-5}). The distribution of the number of BLAST hits per IES (excluding self-hits) is displayed for each species.

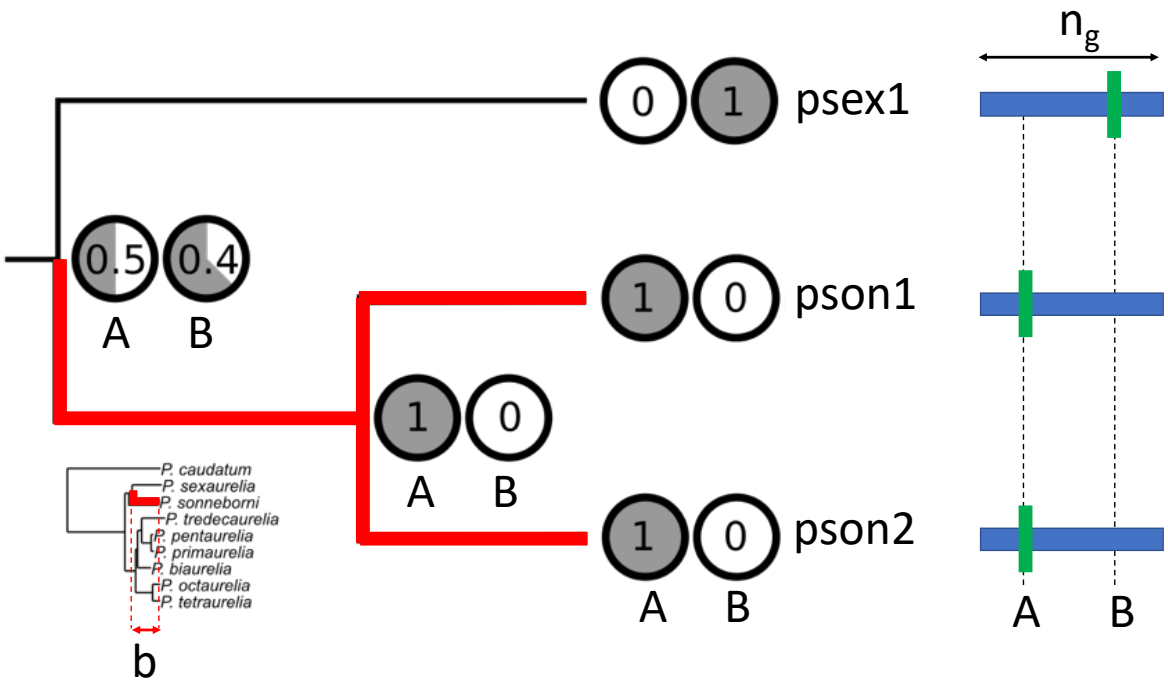


S12 Fig. Estimating the proportion of MIC and MAC DNA in the sample based on IES retention score.

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



S13 Fig. Example of floating IES.
Comparison of MIC and MAC sequences indicates the presence of an IES at this locus. However, because of the presence of a repeated motif at the boundaries of the IES (blue arrows), it is not possible to determine which of the two possible segments (IES-1 in black or IES-2 in red) is actually excised *in vivo*. Such IESs that cannot be unambiguously positioned are called 'floating IES'. They represent 6.8% of the 400,254 IESs detected across all species. In the vast majority of cases (86%) the alternative locations of floating IESs differ by only two bp (as in the example shown here), and there are less than 1% of floating IESs for which the uncertainty in IES position exceeds 5 bp.



S14 Fig. Measuring the rate of IES gain or loss along the species phylogeny.
To illustrate our methodology, we show here an example of a gene family with 3 genes, two from *P. sonneborni* (pson1, pson2) and one from *P. sexaurelia* (psex1). Two IES loci are found in this family (A, B). The probability of presence of an IES (estimated by Bayesian ancestral state reconstruction - see methods) is indicated by shaded circles for each locus at each node of the gene phylogeny. We focus here on the branch of the species tree leading from the common ancestor of *P. sexaurelia* and *P. sonneborni* to the leaf node of *P. sonneborni* (the red branch in the species tree, shown in insert). The length of this branch (*b*), is taken as a proxy for time. Because of a duplication event, this branch of the species tree corresponds to two paths in the gene tree (*k*=2). To estimate the IES gain rate, we calculate for each path the sum of increase in the probability of presence of an IES, for all IES loci (p^+). Along the first path (from the root to pson1), we have p^+A1=0.5 and p^+B1=0. Along the second path (from the root to pson2), we have p^+A2=0.5 and p^+B2=0. The average gain rate along all paths, per unit 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 the number of well aligned sites in the gene family alignment (i.e. the number of sites where the presence of homologous IESs can be assessed). Similarly, to estimate the IES loss rate, we calculate for each path the sum of decrease in the probability of presence of an IES, for all IES loci (p^-). Along the first path (from the root to pson1), we have p^-A1=0 and p^-B1=0.4. Along the second path (from the root to pson2), we have p^-A2=0 and p^-B2=0.4. The average gain rate along all paths, per unit of time and per bp, is thus given by $L=(p^-A1 + p^-B1 + p^-A2 + p^-B2)/(k \times b \times I)$, where *I* is the number IES loci in the gene family (here *I*=2).