Breaking the species barrier: recurrent genomic introgressions from very distant lineages in a ciliate.

Florian Bénitière (1), Olivier Arnaiz (2), Simon Penel (1), Sandra Duharcourt (3), Eric Meyer (4), Linda Sperling (2), Laurent Duret (1)

(1) Universite Claude Bernard Lyon 1, LBBE, UMR 5558, CNRS, VAS, Villeurbanne, France
(2) Université Paris-Saclay, CEA, CNRS - Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette, France
(3) Université Paris Cité, CNRS, Institut Jacques Monod, F-75013 Paris, France
(4) Institut de Biologie de l'ENS (IBENS), Département de Biologie, École Normale Supérieure, CNRS, INSERM, Université PSL, Paris, France

Corresponding Author:

Laurent Duret

Laboratoire Biométrie et Biologie Evolutive,UMR CNRS 5558, Université Lyon 1

43 Bld du 11 Novembre 1918, 69622 Villeurbanne cedex, France

Phone : +33 4 72 44 62 97

Email: Laurent.Duret@univ-lyon1.fr

Author Contributions: FB developed the workflow to detect and classify HT segments. OA and SP were involved in data curation and data analysis. LD, LS and EM performed formal analysis, annotation and data curation. LD, EM, SD and LS were responsible for conception, supervision and funding of the project. LD wrote the manuscript, with the help of EM, SD and LS. All authors approved the final manuscript.

Keywords: Speciation; species barriers; hybridization; ciliates; paramecium

Abstract

The process of speciation, through which two reproductively isolated species emerge from an ancestral population, is often gradual, including a period during which gene flow still occurs, and hence speciation remains reversible. However, if isolation persists, the progressive accumulation of genetic incompatibilities ultimately locks the speciation process, reaching the point where hybrid offspring are inviable or sterile. In plants and animals, the establishment of species barriers is quite rapid: beyond a net genetic divergence of 0.02 substitutions per synonymous site, interspecific introgressions become extremely rare. Here we report that in ciliates, species barriers can be broken despite huge genetic divergence. We show that the genome of *Paramecium sonneborni* contains hundreds of DNA segments that were acquired by recurrent horizontal transfers from multiple very distant lineages (synonymous divergence ~1 substitution/site). These segments approach the size of full-length chromosomes, which indicates that they were introgressed by inter-specific mating. These data show that in the ancestry of *P. sonneborni*, hybridization events with highly divergent species occurred repeatedly and led to viable and fertile offspring. We propose that this unexpected promiscuous sexual behavior is made possible by the fact that paramecia separate somatic and germline functions in two types of nuclei: the elimination of non-self DNA fragments during the development of the somatic genome allows individuals of a species to remain phenotypically the same, despite recurrent genetic flow from distant species into their germline.

Significance Statement

Up to what degree of genetic divergence is it possible to form viable and fertile hybrids between two distinct species? This question is key to identifying the factors contributing to speciation. In plants and animals, the exchange of genetic material between species appears to be extremely rare beyond a divergence of 0.02 substitutions per synonymous site, suggesting that genetic incompatibilities accumulate quite rapidly. By analyzing the genome of the ciliate *Paramecium sonneborni*, we discovered that this unicellular eukaryote recurrently acquired large DNA segments via hybridization with very distantly related species (divergence of up to 1.0 substitutions per site). We propose a mechanistic model to explain how these ciliates can escape the impact of genetic incompatibilities.

Main Text

Introduction

An important issue in evolutionary biology is to understand the process of speciation, i.e. how two reproductively isolated species emerge from an ancestral population. Various factors can initiate the speciation process (e.g. isolation by geographic barriers, adaptation to distinct ecological niches, etc.), but whatever the initial factor, once two populations become reproductively isolated from each other, they start evolving independently, and hence progressively accumulate genetic incompatibilities. These so-called Bateson–Dobzhansky–Muller incompatibilities (BDMIs) decrease the fitness of hybrid individuals, and ultimately lead to their sterility or non-viability (1). The speciation process is often gradual, i.e. includes a 'grey zone' during which genetic introgressions are still possible between "semi-isolated" species (2). Thus, if secondary contacts occur during the early phase, then hybridizations may restore genetic homogeneity, thereby stopping the speciation process. However, as BDMIs accrue, selection is expected to favor the evolution of traits that prevent interspecific mating, thereby further reinforcing genetic isolation. A key point is therefore to unravel the factors that determine the rate at which BDMIs accumulate. The analysis of population genomics data from a wide range of taxa has shown that introgressions are restricted to species that are genetically very close: in animals, gene flow appears to be abolished as soon as the net synonymous divergence exceeds 0.02 substitutions per site, and this

threshold is even lower in plants (3, 4). Thus, although hybridization is quite common in nature, it generally leads to inviable or sterile offspring, and hence does not contribute to gene flow, except between very closely related species (5).

Here we report an extreme exception to this rule, that we discovered incidentally in the course of sequencing the genomes of several paramecia species (6). We present evidence of recurrent events of introgression within the genome of *Paramecium sonneborni*, originating from very remote species (synonymous divergence ~1 substitution/site), and we propose a model explaining how fertile hybrids can be formed between such distant species, despite BDMIs.

To facilitate the interpretation of our observations, it is useful to give some background information on the biology and genetics of *Paramecium* species. These are aquatic unicellular eukaryotes that are commonly found in ponds, where they feed on bacteria, yeast, or algae. They were among the first ciliates to be observed, thanks to the invention of the microscope in the late 17th century, and since then have become a widely used model for biological studies (7). Their life cycle includes a vegetative phase (reproduction by mitotic cell divisions), followed by a sexual phase, which involves either conjugation (the pairing and reciprocal fertilization of two cells of complementary mating types) or autogamy (a self-fertilization process that evolved in an ancestor of the *P. aurelia* clade) (8).

Like all ciliates, *Paramecium* cells possess two types of nuclei: a highly polyploid somatic macronucleus (MAC), responsible for gene expression, and a diploid germline micronucleus (MIC), which is not expressed, but where meiosis occurs. At each sexual generation, the parental MAC is lost, while new MIC and MAC develop from copies of the zygotic nucleus. During MAC development, many DNA segments are reproducibly eliminated, thus leading to a somatic genome that is smaller than the germline genome (9, 10).

The genus *Paramecium* encompasses a wide diversity of species. Based on cytological observations, 19 species can be distinguished (11). However, genetic crosses have revealed a large amount of cryptic diversity within these groups. Notably, early studies have established that *P. aurelia*, initially considered as one 'species', was in fact a complex of 14 sibling species ('syngens') that are morphologically indistinguishable but sexually incompatible (12). More recently, phylogenetic and population genetic studies based on MAC genome sequences have confirmed that these 14 syngens correspond to *bona fide* species, reproductively isolated and genetically divergent (for a review, see (13)). Currently, 20 species have been ascribed to the *P. aurelia* species complex (14): 15 that are morphologically indistinguishable (named *P. primaurelia* to *P. quindecaurelia*), together with *P. sonneborni*, *P. schewiakoffi*, and three sibling species of *P. jenningsi* (12, 15–18). This clade, whose phylogeny is now well established, encompasses a wide variety of species, genetically as distant as the most remotely related mammals (14, 19).

Little is known regarding the ecological factors that may have driven the diversification of this clade. Some species are preferentially found in tropical environments (e.g. *P. sexaurelia, P. octaurelia, P. tredecaurelia, P. quadecaurelia,* and *P. sonneborni*), or in temperate or cold climate zones (e.g. *P. biaurelia* and *P. novaurelia*), whereas others are totally cosmopolitan (e.g. *P. primaurelia* or *P. tetraurelia*) (20). Overall, most species have a very widespread geographical distribution and are found in sympatry on several continents (20). In the *aurelia* clade, the level of neutral polymorphism ranges from 0.6% in *P. tetraurelia* to 2.7% in *P. sexaurelia* and shows no clear pattern of spatial structure, no isolation by distance, indicating recent dispersal (21). One noteworthy observation is that three rounds of whole genome duplications (WGDs) occurred in the ancestry of the *aurelia* lineage, and the last one shortly predates the radiation of the complex (22–24). The differential loss of paralogous gene copies after a WGD is expected to be an important source of BDMIs (22, 25), and hence it has been suggested that these WGDs might have contributed to the diversification of the *P. aurelia* complex (22–24).

To investigate the organization and evolution of the germline genome, we sequenced the MIC of 7 *P. aurelia* species (6, 19). We obtained a high-quality assembly for the *P. tetraurelia* MIC genome (104 Mb), which consists of ~160 tiny chromosomes (300 kb – 1.2 Mb)(6). Three genomic compartments can be distinguished, depending on their fate during the development of the new polyploid MAC: i) MIC-limited regions, that are eliminated from the MAC, ii) constitutive MAC-destined regions, that are systematically retained in the MAC, and iii) MAC-variable regions, which correspond to DNA sequences that are retained in only a small fraction of MAC copies (19). These three compartments represent respectively 20%, 71% and 9% of the MIC genome. The MAC-destined regions are very gene-rich (~500 genes per Mb), whereas most repeated sequences are located in the MIC-limited regions (6, 19). Two types of MIC-limited regions can be distinguished: the 'Internal Eliminated Sequences' (IESs, totalizing 3.9 Mb) that correspond to short sequences (typically <150bp) interspersed within MAC-destined regions and notably within genes, whereas the other eliminated sequences (16 Mb) are generally located towards the end of MIC chromosomes (6). Most other species show similar genome content, with a MIC genome ~100 Mb. The only exception is *P. sonneborni*, where the MIC reaches 217 Mb (6). The size of the MAC-destined compartment is similar across species, ranging from 65 Mb in *P. tredecaurelia* to 82 Mb in *P. sonneborni* (19). Thus, the large size of the *P. sonneborni* MIC results from a huge inflation of its MIC-specific compartment.

To investigate the cause of this increase, we compared *P. sonneborni* MIC sequences to other *Paramecium* genomes, and discovered that *P. sonneborni* contained long regions with very high sequence similarity to distantly related *aurelia* species. This prompted us to systematically investigate patterns of horizontal transfer across paramecia. We found evidence of recurrent events of introgression of chromosome-size segments within the MIC-specific genomic compartment of *P. sonneborni*, originating from very distant species. We propose that these introgressions result from allotetraploidization events, and we discuss the mechanism that could explain how these events can be tolerated despite the accumulation of BDMIs.

Results

Evidence of massive horizontal transfer in the exceptionally large MIC genome of P. sonneborni

To investigate the origin of the large MIC-limited compartment in *P. sonneborni*, we compared its sequences to all available genomes from 13 species of the *P. aurelia* complex (7 species for which both MIC and MAC genomes have been sequenced, and 6 species for which only the MAC is available). *P. sonneborni* is closely related to *P. jenningsi* (genetic distances at synonymous sites: dS=0.27), but quite distant from all the other species that we analyzed (Figure 1A), with dS values ranging from 0.8 to 1.0 substitution per site (19). This implies that outside of coding regions, sequences are generally too divergent to be aligned across these species. Yet, we noticed that several *P. sonneborni* MIC contigs contain long segments with very high similarity to distantly related species. For instance, we observed a *P. sonneborni* contig of 100kb presenting long stretches above 99% identity with a MAC scaffold from *P. quadecaurelia*, a level of similarity higher than with any other *aurelia* species, including *P. jenningsi* (Figure 1B). This suggests that this long genomic segment has been subjected to horizontal transfer (HT) between the *P. sonneborni* and *P. quadecaurelia* lineages.

This unexpected observation led us to systematically analyze all MIC genomes (from 7 species) to search for loci presenting evidence of recent HT among the *aurelia* complex. For this, we divided each MIC genome (query) into non-overlapping 500 nt windows and compared (with blastn) each window with all other available *Paramecium* genomes (targets) (Figure 1A). We then searched for sets of neighboring windows (at least 2 kb-long in total) with conserved order and orientation between the query and the target, and showing a strong signal of HT (i.e. a level of sequence similarity significantly higher with a genome that is not the sister group of the focal species, than with any other *aurelia* species).

150 In most species we identified only a limited number of HT segments (1 to 22 segments, totalizing
151 up to 144 kb; Figure 1C). Strikingly, in the *P. sonneborni* MIC genome, we identified 521 HT
152 segments totaling almost 10 Mb. HT segments are on average 2 to 5 times longer in *P. sonneborni*
153 (mean length = 18.4 kb) than in other species (Suppl. Fig. S1). In *P. sonneborni* we detected 120
154 HT segments longer than 25 kb (totaling 5 Mb). The length of HT segments in *P. sonneborni* is
155 probably underestimated because the MIC assembly is quite fragmented (N50 ~33 kb). Indeed, we
156 found several cases where a single region in the MAC genome of a target species showed synteny
157 with multiple *P. sonneborni* contigs (Figure 1B). These observations indicate that many HT
158 segments in *P. sonneborni* are very long, approaching the size of an entire chromosome (>220 kb
159 for the example shown in Figure 1B). Conversely, none of the HT segments in the other species
160 were larger than 25 kb.

161 Several observations allowed us to exclude the hypothesis of a DNA contamination artefact. First,
162 contamination with DNA from the target genomes should lead to nearly identical sequences,
163 whereas for many HT segments, the level of sequence divergence ranges between 1% and 12%
164 (Figure 2), far too high to result from sequencing errors. Second, two of the target genomes with
165 which we detected many HT segments (*P. quadecaurelia* and *P. novaurelia;* Figure 1A) were
166 sequenced in the USA (24), and were never handled in the laboratory where *P. sonneborni* MAC
167 and MIC DNA were sequenced (Genoscope, France). Finally, the sequencing depth along *P.*
168 *sonneborni* HT segments is the same as that of the rest of the genome, which would not be
169 expected if they resulted from contaminant DNA (Suppl. Fig. S2).

170 The HT segments detected in *P. sonneborni* involve four distinct lineages (represented by *P.*
171 *sexaurelia, P. tredecaurelia, P. quadecaurelia,* and *P. novaurelia*) (Figure 1A). Among HT
172 segments detected by comparison between *P. sonneborni* and *P. sexaurelia* (the only species pair
173 for which we had both MIC genome assemblies), we observed many conserved IESs (Suppl. Fig.
174 S3), which demonstrates that these HT events involved the transfer of MIC DNA and not of MAC
175 DNA (where IESs are absent). *P. sexaurelia* belongs to the same subclade as *P. sonneborni* but is
176 nevertheless very distant (average dS=0.8; (19)). The three other lineages belong to subclade A,
177 very divergent from *P. sonneborni* (dS~1; (19)). We did not detect any HT segment with the other
178 lineages of subclade A (Figure 1A).

179 It should be noted that MIC genome sequences are available for only 7 of the 13 analyzed species.
180 Because of this sampling limitation, HT segments present in the MIC genome of the query were
181 detectable only if they encompassed MAC-destined regions in the target genome (see Methods for
182 details). Hence, the set of segments that we identified is probably an underestimate of the total
183 amount of HT.

184 **P. sonneborni is the recipient of recurrent horizontal transfers from distantly related**
185 **lineages**
186 The above results show that gene flow occurred between *P. sonneborni* and several other
187 *Paramecium* lineages. To establish the direction of these transfers, we analyzed the distribution of
188 HT segments in the MIC of *P. sonneborni*: 85% are located in the MIC-limited compartment of *P.*
189 *sonneborni*, 14% in MAC-variable regions and only 1% in MAC-destined regions (these
190 compartments represent respectively 55%, 7% and 38% of the MIC genome) (Figure 1A). Thus,
191 the majority of HT segments are confined to genomic compartments that are either eliminated or
192 variably retained during MAC development. As a result, although the HT segments correspond to
193 MAC-destined regions in the target genomes and are therefore very gene-rich, the genes they
194 contain are unlikely to be expressed in *P. sonneborni*. In total, among the 521 HT segments, we
195 identified only 35 genes located in the MAC-destined compartment of *P. sonneborni* (Suppl. Table
196 S1). Comparison with their homologs in other *aurelia* species showed that at least 25 of them are
197 pseudogenes (containing nonsense or frameshift mutations), and among the ten that appear to be
198 intact, only four show evidence of transcription (> 1RPKM) in vegetative cells. Thus, even though
199 HT segments detected in *P. sonneborni* included several thousand genes, most of them have lost

their functionality. This implies that *P. sonneborni* was the recipient, and not the donor of the horizontal transfers.

We did not detect any case of a gene from the donor genome integrated by recombination at its homologous locus in the *P. sonneborni* MIC genome. Most of the HT segments are located in the MIC-limited compartment, and hence away from genic regions, and the few (pseudo)genes that we detected in the MAC-destined compartment are most often located towards the end of MAC chromosomes (Suppl. Table S1). Thus, these HT events did not lead to the replacement of recipient genes by their donor homologues, but to the addition of new gene copies (i.e. gene duplication via HT).

P. sonneborni has received HT from four distinct donor lineages (Figure 1A). Several observations indicate that the detected HT segments result from numerous independent genetic transfers, and not from just four occasional events. Indeed, the level of sequence divergence observed between *P. sonneborni* HT segments and the target genomes varies widely among HT segments, ranging from ~0% to ~12%, even among segments coming from the same lineage (Figure 2). Two non-exclusive hypotheses can be proposed to explain this large variance: i) some HT events are more ancient than others, and hence have had more time to accumulate divergence between the donor and the recipient lineages, or ii) HT segments have about the same age, but originated from multiple related lineages that vary in their level of divergence from the reference species used in our analyses. One noteworthy observation is that the length of HT segments tends to decrease as the sequence divergence increases (Figure 2): for a given donor lineage, the longest HT segments (approaching the size of full-length chromosomes) are generally less divergent from the reference target than other HT segments, whereas the most divergent HT segments are always relatively short. The most parsimonious explanation for this pattern is that after the transfer of full-length chromosomes, HT segments progressively shorten via chromosomal rearrangements (inversions, translocations, deletions) and concomitantly diverge via point substitutions. This therefore suggests that the variance in level of divergence mainly reflects the variability in the age of HT events. In any case, these observations imply that HT events from distant species occurred recurrently in the lineage of *P. sonneborni.*

Little is known regarding the ecology of *P. sonneborni*. Although this species is rare, it is distributed across very wide geographical areas (the three strains reported so far are from Texas, Dominican republic and Cyprus) (15, 26, 27). The high frequency of introgression coming from *P. sexaurelia, P. tredecaurelia*, and *P. quadecaurelia* lineages might be due to the fact that like *P. sonneborni*, these species are preferentially found in warm climate zones (20).

In principle, when two species are able to form hybrids, introgressions may occur in both directions. Unfortunately, we could not assess the presence of HT segments related to *P. sonneborni* in any of the four lineages from which it received DNA (represented in our data set by *P. tredecaurelia, P. quadecaurelia, P. novaurelia* and *P. sexaurelia*): for the first three, we do not have a MIC genome assembly, and for *P. sexaurelia*, we do not have an appropriate sister group to detect HT from *P. sonneborni*. Thus, the question of whether transfers occur in both directions remains to be investigated.

Discussion

Recurrent hybridization with genetically highly divergent species.

Hybridization between closely related species is quite common in nature, but often leads to inviable or sterile offspring (5). Different types of processes cause this decreased fitness of hybrids. First, after their divergence, sister species tend to accumulate genetic incompatibilities (i.e. alleles that are neutral or beneficial in one species but that interact negatively with the alleles in the other species), so-called Bateson–Dobzhansky–Muller incompatibilities (BDMIs) (1). Notably, as each species has adapted to its specific ecological niche, the mixed set of alleles in the hybrids may turn

249 out to be unfit in both environments. Second, chromosomal rearrangements and sequence
250 divergence impair homologous recombination, and thereby cause incorrect chromosome
251 segregation during meiosis in hybrids (28). Hence, the probability of forming fertile hybrids (and the
252 potential of inter-species gene flow) decreases with increasing genetic distance (5). Typically, in
253 animals and plants, the amount of gene flow across populations/species drops drastically beyond
254 2% of net synonymous divergence (3, 4) and in budding yeast, hybrid viability falls to almost zero
255 above 10% of genetic divergence (28).

256 Here we show that a large fraction of the *P. sonneborni* genome has been acquired via recurrent
257 events of horizontal transfer (HT) from multiple lineages. Cases of horizontal transfers had already
258 been reported in the *P. aurelia* complex (19), but they were of relatively short size (<3.5kb) and
259 corresponded to transposable elements. Such transfers might be mediated by vectors (e.g.
260 viruses), and therefore do not necessarily involve inter-specific mating. The size of the longest HT
261 segments detected in *P. sonneborni* (> 200 kb), is in the range of germline chromosome sizes (300
262 kb to 1 Mb) (6). This implies that they were most probably transferred as full-length chromosomes,
263 through inter-specific conjugation. Thus, our observations reveal that the *P. sonneborni* lineage
264 recurrently mated and produced fertile offspring with very distant species (synonymous divergence:
265 dS~0.8 to 1 substitution per site). For comparison, this corresponds to the genetic distance between
266 the most divergent mammalian taxa (e.g. monotremes vs. placental mammals (29)).

267 Known examples of distant species able to produce viable hybrid offspring are quite rare. The three
268 most extreme cases reported so far are two species of fern that diverged ~60 Mya (30), sturgeon
269 and paddlefish that diverged ~184 Mya (31), and two gar fish species that diverged ~100 Mya (32).
270 The latter case is particularly striking, because unlike the other two, the hybrids are not only viable,
271 but also fertile (32). It has been suggested that the absence of strong BDMI in gars results from the
272 fact that their genomes evolve extremely slowly (32). The same explanation probably holds for the
273 two other cases: the median dS between sturgeon and paddlefish is only 0.12 (33), and in the fern,
274 the dS between the two parental genomes is 0.1 (30). Thus, even though these species have been
275 separated for a very long period, they are genetically quite close. Although introgressions are rare
276 at this level of genetic divergence (3, 4), several cases have been reported. For instance, gene flow
277 has been observed in the *Drosophila yakuba* clade (~3 Mya of separation), or in *Ciona* (sea squirts;
278 1.5-2 Mya of separation), between species showing a dS of 0.12 to 0.14 (34, 35). But to our
279 knowledge, *P. sonneborni* is the first reported case of a eukaryote showing promiscuous sexual
280 behavior with genetically highly diverged species (dS~1).

281

282 **Remaining the same despite hybridization with distant species.**
283 Our observations raise three main questions. First, it is unclear how such distant species could
284 have successfully interbred. In standard lab conditions, *P. sonneborni* was not observed to
285 conjugate with any other *P. aurelia* species, not even with the closely related *P. jenningsi* (15, 26).
286 More extensive analyses in other species concluded that inter-specific conjugation may be blocked
287 at a number of different steps: mating type recognition and agglutination, formation of conjugating
288 pairs, cross-fertilization, and development of functional zygotic MACs. Although all *P. aurelia*
289 species have two homologous mating types (O and E), inter-specific agglutination of O and E cells
290 is only observed in some pairs of species from the same subclade, and in most cases it does not
291 proceed to the formation of conjugating pairs (12). Inter-specific transformation experiments have
292 shown that the E-specific transmembrane protein mtA is directly responsible for the specific
293 recognition of O cells of the same species (36). In the lab, the need for mating type recognition can
294 be bypassed by chemical treatments that induce the successful conjugation of a single mating type
295 with itself (37). Chemical treatments can also induce the formation of conjugating pairs between
296 species (e.g. between a *P. aurelia* species and the outgroup *P. multimicronucleatum*), but in such
297 cases genetic exchange appears to fail and surviving ex-conjugants likely result from self-
298 fertilization (37). Even in the best of cases where inter-specific conjugation can occur naturally, e.g.

between the closely related *P. tetraurelia* and *P. octaurelia*, most ex-conjugants do not survive, genetic exchange rarely succeeds, and the rare viable true hybrids are all completely sterile (38).

While the origin of the *P. sonneborni* hybrid ancestors thus remains mysterious, a second question is how they could escape the negative consequences of BDMIs. Here the answer may lie in the peculiar nuclear dimorphism of ciliates. Indeed, the HT segments identified in the genome of *P. sonneborni* are essentially confined to its MIC-limited compartment and hence are not expressed, which effectively hides genetic incompatibilities. In *P. aurelia* species, the distinction between segments that are retained or eliminated during MAC development relies on a comparison between the maternal MIC and the maternal MAC, via small RNAs that are produced from the entire MIC genome during meiosis and are still present in the cytoplasm of the zygote after conjugation: the so-called scnRNAs mediate a genomic subtraction that targets the elimination from the new developing MAC of all maternal MIC sequences that are absent from the maternal MAC (9, 10) (Figure 3). We conjecture that in *P. sonneborni*, a similar mechanism also leads to the elimination of sequences that are present in the paternal MIC but absent from the maternal MAC. This would explain how, after conjugation of *P. sonneborni* with a distantly related species, the chromosomes from the latter may be systematically eliminated from the developing MAC in the F1 hybrid carrying the *P. sonneborni* cytoplasm. That the current species indeed derives from that side of the cross after each inter-specific mating is indicated by the *P. sonneborni* mitochondrial genome (39), whose phylogenetic position is in total agreement with the species tree inferred from MAC nuclear genomes (14) (Figure 3). In other words, the successful matings of *P. sonneborni* ancestors with distant relatives may have been facilitated by the capacity of *P. aurelia* species to exclude non-self DNA from the zygotic MAC (self DNA being defined by the maternal MAC genome). Interestingly, a similar situation has been reported in a dipteran insect, which possesses two chromosomes that have been acquired by introgression from a distant lineage, but that are restricted to the germline, and hence are not expressed in the soma (40).

The third question raised is how the F1 hybrid MIC may have managed to complete meiosis to perpetuate itself. Indeed, the most likely cause of sterility in true F1 hybrids is meiotic failure, due to sequence divergence and/or karyotypic differences between the two parental genomes. We propose that in the F1 hybrid ancestors of the current *P. sonneborni*, fertility may have been restored by whole-genome duplications in the hybrid MIC, resulting in allotetraploid F1s in which the two sets of duplicated chromosomes can be readily distinguished by the meiotic recombination machinery - thereby allowing proper segregation during meiosis. These fertile allotetraploids would retain the original *P. sonneborni* phenotype essentially unchanged, however, since introgressed genes cannot be expressed. In the absence of any selective pressure to keep them functional, introgressed genes would rapidly decay into pseudogenes and, ultimately, into unrecognizable genome “dark matter”.

Although our model explains how *P. aurelia* species might escape the deleterious consequences of inter-specific mating, it leaves open further questions. First, it does not provide any clue regarding the *raison d’être* of these recurrent introgressions. Because introgressed genes can hardly be expressed, these events do not appear to contribute much to the adaptive potential of *P. sonneborni*. Second, this model predicts that inter-specific mating can be tolerated only if it involves a species that is genetically distant enough to be recognized as non-self during MAC development and meiosis (Figure 3). Hence, selection should still favor the evolution of traits that prevent inter-specific pairing. A third intriguing point is the fact that this promiscuous behavior is observed only in the *P. sonneborni* lineage. For *P. sonneborni* and *P. sexaurelia*, our experimental design provided the same power to detect HT events originating from subclade A. Yet, whereas we identified ~6,000 kb of HT segments coming from subclade A in *P. sonneborni,* we detected only 76 kb in *P. sexaurelia* (Fig. 1A,C). Moreover, we detected very few HT segments in the five other MIC genomes analyzed, and they are all short, suggesting that they result from a different mode of horizontal transfer (e.g. via viral vectors) or that they result from much more ancient inter-specific

349 hybridization. So far, *P. sonneborni* appears to be unique among paramecia in showing such a high
350 rate of introgression in its recent ancestry.

351 We speculate that the large size of the MIC-limited compartment in *P. sonneborni* (135 Mb
352 compared to ~30 Mb in other *aurelia* species) might result from its propensity to hybridize with
353 distant species. Although the HT segments that we identified in *P. sonneborni* cover only 7.4% of
354 its MIC-limited compartment, they probably represent only the visible tip of the iceberg. Indeed, our
355 study design allowed the detection of HT segments only if they encompassed MAC-destined
356 regions from the donor genome and if they were closely related (>80% sequence identity) to the
357 available reference genomes. Given that sequences in the MIC-limited compartment cannot be
358 expressed, they diverge rapidly, like pseudogenes. It is therefore plausible that a large fraction of
359 the MIC-limited compartment of *P. sonneborni* corresponds to ancient HT segments that are no
360 longer recognizable. Thus, the large size of the genome of *P. sonneborni* might simply reflect the
361 recurrent gains of full sets of chromosomes via allotetraploidization, followed by progressive
362 shrinking through deletions or chromosome loss.

363 The genome of *P. sonneborni* provides an interesting perspective on the definition of species. The
364 model we propose (Figure 3) explains how it is possible for a species to persist, despite recurrent
365 hybridization with very distant lineages. Indeed, owing to nuclear dimorphism, hybrids remain
366 phenotypically identical to their maternal lineage, even if they have received a full set of
367 chromosomes from a very divergent species. Moreover, they also remain interfertile with their
368 parental lineage. Thus, as long as gene flow is maintained in the MAC-destined compartment
369 (which represents only ~35% of the *P. sonneborni* genome), individuals can be considered as being
370 part of the same species, even if their MIC-limited compartment is not shared. This tentative
371 extension of the definition of a species echoes Sonneborn's own hesitation in recognizing that what
372 he initially called "varieties" of *P. aurelia* (8) are actually sexually incompatible species (12).
373 Ironically, the promiscuous sibling is the one that bears the name of the father of *Paramecium*
374 genetics.

375

376 **Materials and Methods**

377 **Detection of horizontal transfer across aurelia species**

378 We analyzed the MIC genome of each species *(P. tetraurelia, P. octaurelia, P. biaurelia, P.*
379 *pentaurelia, P. primaurelia, P. sexaurelia* and *P. sonneborni*) (6) to search for segments that have
380 been subject to horizontal transfer (HT) across the *aurelia* complex (see Suppl. Table S2 for a full
381 list of strains and genome assemblies used in this study). For this, the MIC genome of each focal
382 species was cut into 500 bp-long windows, that were compared to all genomes (MAC and MIC, if
383 available) of the 12 other aurelia species with BLASTN (version 2.5.0+; (41)) using the following
384 parameters:

385 `blastn -query <infile> -db <db> -out <File_Out> -task blastn -evalue 1e-6 -`
386 `max_target_seqs 1`

387 For each window, we retained the best blast hit on each of the target genomes. We then selected
388 windows having their best hit with a species that is not a sister-group of the focal species (see
389 phylogenetic tree in Figure 1A). We retained windows with a high level of sequence similarity (>80%
390 identity), significantly higher than with any other species (> mean + 2 standard deviations). These
391 stringent criteria were chosen to identify windows with strong evidence of recent HT.

392 Then, starting from these 'seed windows', we extended the HT segment by aggregating neighbor
393 windows (< 5kb apart) having their best hit on the same scaffold of the non-sister species. The
394 extension was stopped as soon as a seed window matching another scaffold was reached or when
395 the distance to the next window matching the same scaffold exceeded 5 kb. Finally, we retained
396 HT segments that contained at least 4 seed windows, and that included at least 70% of windows

having their best hits on the same scaffold. The threshold of 4 seed windows was chosen to ensure that HT segments are at least 2kb-long.

We first performed the HT search by comparing MIC windows of the focal species (query) against the MAC genomes of the 12 other *aurelia* species (targets). We then repeated the analyses using as targets the MIC instead of the MAC genomes, for the 7 species for which both the MAC and the MIC were available. In *P. sonneborni*, this second strategy identified 275 additional candidate HT segments, all matching with the MIC of *P. sexaurelia*. However, given that we only have the MAC but not the MIC genome of *P. jenningsi*, we could not ascertain that these segments were absent from the sister-group. In the other species, both strategies gave very similar results (with at most 8 new HT segments identified). To be conservative, the analyses presented in the main text were restricted to HT segments detected with the first strategy (MAC targets).

The median synonymous divergence (dS) between *P. jenningsi* and *P. sonneborni* was computed on the set of 473 one-to-one orthologs reported in a previous study (14).

Analysis of HT genes located in the MAC-destined compartment

Based on annotations of the *P. sonneborni* MAC genome assembly (24), we identified 39 gene predictions located in HT segments encompassing MAC-destined regions. For each of them, we used BLASTP to identify homologs in other species of the *P. aurelia* complex. We computed multiple alignments to investigate the conservation of each protein family. In 36 out of 39 cases, we identified nonsense and/or frameshift mutations in the *P. sonneborni* predicted gene with respect to its homologs. We also analyzed the expression level of these predicted genes using RNAseq data from *P. sonneborni* vegetative cell cultures (14).

Data availability

All genome and annotation files are available from the NCBI or from the ParameciumDB download section (links and accession numbers are provided in Supplementary Table S2). The complete list of detected HT segments and R scripts required to reproduce the figures are available on Zenodo (https://doi.org/10.5281/zenodo.18684075).

Acknowledgments

This work was performed using the computing facilities of the CC LBBE/PRABI. We thank Carina Mugal and Mark Stoneking for helpful comments on the manuscript. This work was supported by the Centre National de la Recherche Scientifique, and by the Agence Nationale de la Recherche (ANR-18-CE12-0005/LaMarque). Work in SD laboratory is supported by the Agence Nationale de la Recherche (ANR-23-CE12-0027; ANR-25-CE12-7757), and by the Fondation de la Recherche Medicale (https://www.frm.org)(Equipe FRM EQU202203014643).

References

1. H. A. Orr, Dobzhansky, Bateson, and the Genetics of Speciation. *Genetics* **144**, 1331–1335 (1996).

2. S. Stankowski, M. Ravinet, Defining the speciation continuum. *Evolution* **75**, 1256–1273 (2021).

3. C. Roux, *et al.*, Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. *PLOS Biol.* **14**, e2000234 (2016).

4. F. Monnet, *et al.*, Rapid establishment of species barriers in plants compared with that in animals. *Science* **389**, 1147–1150 (2025).

5. N. B. Edelman, J. Mallet, Prevalence and Adaptive Impact of Introgression. *Annu. Rev. Genet.* **55**, 265–283 (2021).

6. O. Arnaiz, *et al.*, The tiny germline chromosomes of Paramecium aurelia have an exceptionally high recombination rate and are capped by a new class of Helitrons. [Preprint] (2025). Available at: https://www.biorxiv.org/content/10.1101/2025.11.06.686955v1 [Accessed 3 February 2026].

7. J. Beisson, *et al.*, Paramecium tetraurelia: The Renaissance of an Early Unicellular Model. *Cold Spring Harb. Protoc.* **2010**, pdb.emo140-pdb.emo140 (2010).

8. T. M. Sonneborn, Sex, Sex Inheritance and Sex Determination in Paramecium Aurelia. *Proc. Natl. Acad. Sci. U. S. A.* **23**, 378–385 (1937).

9. M. Bétermier, L. A. Klobutcher, E. Orias, Programmed chromosome fragmentation in ciliated protozoa: multiple means to chromosome ends. *Microbiol. Mol. Biol. Rev.* **0**, e00184-22 (2023).

10. T. Balan, L. K. Lerner, D. Holoch, S. Duharcourt, Small-RNA-guided histone modifications and somatic genome elimination in ciliates. *WIREs RNA* **15**, e1848 (2024).

11. B. A. De Souza, R. J. P. Dias, M. V. X. Senra, Intrageneric evolutionary timing and hidden genetic diversity of Paramecium lineages (Ciliophora: Oligohymenophorea). *Syst. Biodivers.* **18**, 662–674 (2020).

12. T. Sonneborn, The Paramecium aurelia complex of 14 sibling species. *Trans Am Microsc Soc* **94**, 155–178 (1975).

13. H. Long, *et al.*, Paramecium Genetics, Genomics, and Evolution. *Annu. Rev. Genet.* **57**, 391–410 (2023).

14. N. Sawka-Gądek, *et al.*, Evolutionary Plasticity of Mating-Type Determination Mechanisms in Paramecium aurelia Sibling Species. *Genome Biol. Evol.* **13** (2021).

15. K. J. Aufderheide, P.-M. Daggett, T. A. Nerad, Paramecium sonneborni n. sp., a New Member of the Paramecium aurelia Species-Complex1. *J. Protozool.* **30**, 128–131 (1983).

16. S. I. Fokin, *et al.*, Morphological and molecular investigations of Paramecium schewiakoffi sp. nov. (Ciliophora, Oligohymenophorea) and current status of distribution and taxonomy of Paramecium spp. *Eur. J. Protistol.* **40**, 225–243 (2004).

17. E. PrzyboŚ, S. Tarcz, Global molecular variation of Paramecium jenningsi complex (Ciliophora, Protista): a starting point for further, detailed biogeography surveys. *Syst. Biodivers.* **17**, 527–539 (2019).

18. A. Potekhin, R. Mayén-Estrada, Paramecium Diversity and a New Member of the Paramecium aurelia Species Complex Described from Mexico. *Diversity* **12**, 197 (2020).

19. D. Sellis, *et al.*, Massive colonization of protein-coding exons by selfish genetic elements in Paramecium germline genomes. *PLOS Biol.* **19**, e3001309 (2021).

20. M. Melekhin, A. Potekhin, E. Gentekaki, C. Chantangsi, Paramecium (Oligohymenophorea, Ciliophora) diversity in Thailand sheds light on the genus biogeography and reveals new phylogenetic lineages. *J. Eukaryot. Microbiol.* **71**, e13004 (2024).

21. P. Johri, *et al.*, Population Genomics of Paramecium Species. *Mol. Biol. Evol.* **34**, 1194–1216 (2017).

22. J.-M. Aury, *et al.*, Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* **444**, 171–178 (2006).

23. C. L. McGrath, J.-F. Gout, P. Johri, T. G. Doak, M. Lynch, Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Res.* **24**, 1665–1675 (2014).

24. J.-F. Gout, *et al.*, Dynamics of Gene Loss following Ancient Whole-Genome Duplication in the Cryptic Paramecium Complex. *Mol. Biol. Evol.* **40**, msad107 (2023).

25. D. R. Scannell, K. P. Byrne, J. L. Gordon, S. Wong, K. H. Wolfe, Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**, 341–345 (2006).

26. E. Przyboś, S. Tarcz, M. Rautian, N. Lebedeva, The first European stand of *Paramecium sonneborni* (*P. aurelia* complex), a species known only from North America (Texas, USA). *Eur. J. Protistol.* **50**, 236–247 (2014).

27. S. Tarcz, M. Surmacz, E. Przyboś, Sampling hidden microbial eukaryotic biodiversity in the tropics: new insights from the Paramecium aurelia complex (Ciliophora, Protozoa). *Folia Biol. (Praha)* **71**, 159–170 (2023).

28. G. O. Bozdag, J. Ono, Evolution and molecular bases of reproductive isolation. *Curr. Opin. Genet. Dev.* **76**, 101952 (2022).

29. W. C. Warren, *et al.*, Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**, 175–183 (2008).

30. C. J. Rothfels, *et al.*, Natural Hybridization between Genera That Diverged from Each Other Approximately 60 Million Years Ago. *Am. Nat.* **185**, 433–442 (2015).

31. J. Káldy, *et al.*, Hybridization of Russian Sturgeon (Acipenser gueldenstaedtii, Brandt and Ratzeberg, 1833) and American Paddlefish (Polyodon spathula, Walbaum 1792) and Evaluation of Their Progeny. *Genes* **11**, 753 (2020).

32. C. D. Brownstein, *et al.*, The genomic signatures of evolutionary stasis. *Evolution* **78**, 821–834 (2024).

33. A. K. Redmond, D. Casey, M. K. Gundappa, D. J. Macqueen, A. McLysaght, Independent rediploidization masks shared whole genome duplication in the sturgeon-paddlefish ancestor. *Nat. Commun.* **14**, 2879 (2023).

34. D. A. Turissini, D. R. Matute, Fine scale mapping of genomic introgressions within the Drosophila yakuba clade. *PLOS Genet.* **13**, e1006971 (2017).

516 35. C. Fraïsse, *et al.*, Introgression between highly divergent sea squirt genomes: an adaptive
517 breakthrough? *Peer Community J.* **2** (2022).

518 36. D. P. Singh, *et al.*, Genome-defence small RNAs exapted for epigenetic mating-type
519 inheritance. *Nature* **509**, 447–452 (2014).

520 37. A. Miyake, Induction of conjugation by chemical agents in Paramecium. *J. Exp. Zool.* **167**,
521 359–379 (1968).

522 38. B. W. Haggard, Interspecies Crosses in Paramecium aurelia (Syngen 4 by Syngen 8). *J.*
523 *Protozool.* **21**, 152–159 (1974).

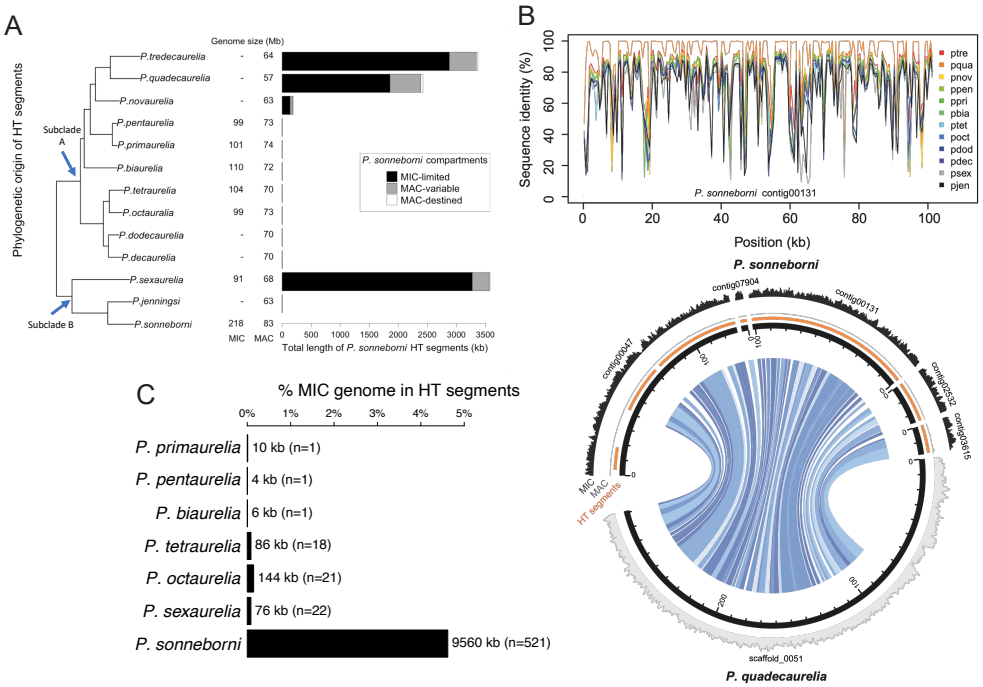
524 39. O. Arnaiz, E. Meyer, L. Sperling, ParameciumDB 2019: integrating genomic data across the
525 genus for functional and evolutionary biology. *Nucleic Acids Res.* **48**, D599–D605 (2020).

526 40. C. N. Hodson, K. S. Jaron, S. Gerbi, L. Ross, Gene-rich germline-restricted chromosomes in
527 black-winged fungus gnats evolved through hybridization. *PLOS Biol.* **20**, e3001559 (2022).

528 41. C. Camacho, *et al.*, BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
529 (2009).

530

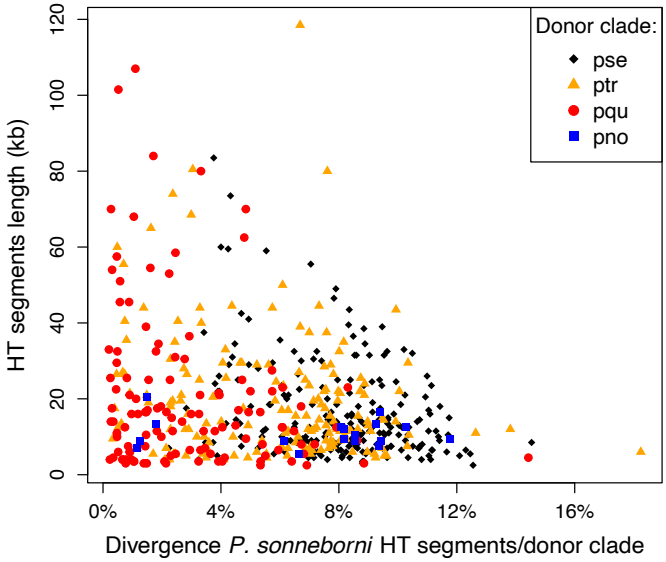
531 **Figure 1.**



532
533 **Figure 1. Detection of horizontally transferred DNA segments in Paramecia. A.** Phylogenetic tree of
534 *Paramecium aurelia* species (modified from (24)). For each species, the sizes of the MAC and, if available,
535 MIC genome assemblies are indicated. All the *aurelia* species have MAC genomes around 70 Mb and MIC
536 genomes around 100 Mb in size, with the notable exception of *P. sonneborni* whose MIC genome is more
537 than twice as large (6). The barplot indicates the phylogenetic origin of HT segments identified in the *P.*
538 *sonneborni* genome. HT regions can be localized in three genomic compartments: MIC-limited (black), MAC-
539 variable (dark grey) and MAC-destined (light grey). **B.** Example of a region of the *P. sonneborni* MIC genome
540 that presents the hallmarks of HT from other *P. aurelia* species. The upper panel presents the data for *P.*
541 *sonneborni* contig00131: the contig was cut into 500bp-long non-overlapping windows, and each window was
542 compared (BLASTN) to all other *P. aurelia* MAC genomes. Each colored line indicates the level of sequence
543 identity with their best hit in the corresponding target genome. HT segments were identified in *P. sonneborni*
544 by searching for sets of neighboring windows with conserved synteny, and whose best hits were not with its
545 sister group (*P. jenningsi*, pjen) (see Methods for details). In this example, the entire *P. sonneborni* contig
546 shows a very high level of similarity with the genome of *P. tredecaurelia* (ptre), more than with any other *P.*
547 *aurelia* species. Species codes are given in Supplemental Table S2. Lower panel, the circos plot shows the
548 conservation of synteny between five *P. sonneborni* contigs containing HT segments (orange) and the putative
549 donor region of the *P. quadecaurelia* MAC genome (scaffold_0051). Histograms describe DNA-seq coverage
550 for MIC DNA sample (black) and MAC DNA sample (grey). In this example, the HT segments are located in
551 the MIC-limited compartment of *P. sonneborni*: they are well covered in the MIC DNA sample but not in the
552 MAC. The colors of the arcs are defined by the similarity between regions. **C.** Percentage of MIC genomes in
553 horizontally transferred (HT) segments. The numbers next to each bar indicate the cumulative size and
554 number of HT segments.

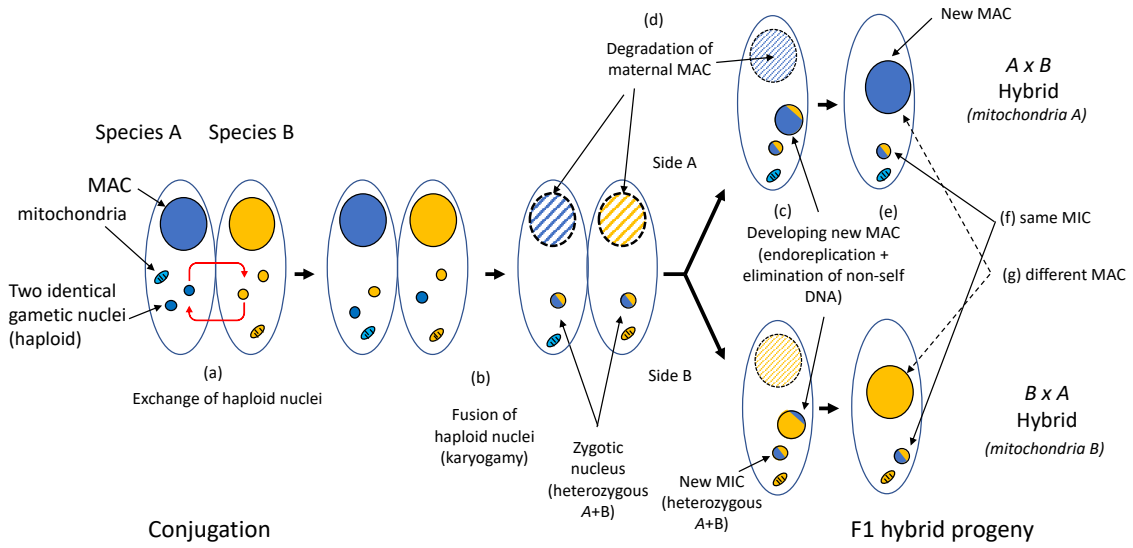
555

556 **Figure 2.**
557
558
559
560



561
562 **Figure 2. Relationship between the size of HT segments in the *P. sonneborni* genome, and their**
563 **divergence from the donor lineage.** The origin of the donor clade is indicated by different colors and dot
564 styles. Species codes are given in Supplemental Table S2.
565
566
567

568 **Figure 3.**
569



570
571 **Figure 3. Model: how the nuclear dimorphism in Paramecia might hide interspecific genetic**
572 **incompatibilities**. Sexuality in paramecia occurs through conjugation: cells with compatible mating types pair
573 and engage meiosis. Three of the four resulting haploid nuclei are eliminated, and the remaining one
574 undergoes a mitotic division, leading to the formation of two identical haploid gametic nuclei. The two paired
575 cells exchange one haploid nucleus (a), and then, in each cell, the haploid nuclei fuse to form the new zygotic
576 nucleus (b). After two mitotic divisions, one nucleus differentiates to form the new MAC, while the other
577 constitutes the new MIC (c) (for the sake of simplicity, only one MIC nucleus is shown in the figure; see (Zhang
578 et al., 2023) for more details), while the maternal MAC nucleus is progressively degraded (d). During the
579 development of the new MAC, its chromosomes are amplified by endoreplication, and a fraction of the genome
580 (termed MIC-limited) is eliminated. The distinction between segments that are retained or eliminated relies on
581 a comparison between the maternal MIC and the maternal MAC, via small RNAs produced during meiosis:
582 sequences that are present in the maternal MIC but absent from the maternal MAC are eliminated (c) (Balan
583 et al., 2024). Hence, even though the two F1 progeny have exactly the same germline genotype (f), their MAC
584 content may differ, depending on the content of their respective maternal MACs (g). The model that we
585 propose relies on the assumption that during interspecific crosses, sequences that are present in the paternal
586 MIC but absent from the maternal MAC are also eliminated. This would explain how crosses between distant
587 species can lead to a viable and fertile offspring. Indeed, let us consider a conjugation event between two
588 genetically divergent species (say A and B), and let us focus on the daughter cell from the A side of the cross
589 (a). Its new MIC is allotetraploid, with the full sets of chromosomes from A and B (b). But given that B
590 sequences are divergent from the A sequences present in its maternal MAC, they will all be eliminated during
591 the development of its new MAC (c). Thus, in the side A daughter cell, the new MAC only expresses genes
592 coming from the A genome, thereby effectively hiding any BDMI with the B genome. Hence, although
593 genetically hybrid (A+B), the side A daughter cell should be viable and retain the phenotype of its maternal
594 lineage (A). Moreover, these cells should be able to complete meiosis normally, as the divergence between
595 the two genomes limits the risk of mispairing between homeologous A/B chromosomes. In F2 individuals
596 resulting from a backcross between an AxB hybrid and an A individual, meiosis is expected to proceed
597 normally for A chromosomes, and hence, even though some of their B chromosomes mis-segregate, these
598 F2's should lead to a viable and fertile offspring. This model can account for the introgression of DNA coming
599 from B in the MIC-limited compartment of species A.

Supplementary Information

Breaking the species barrier: recurrent genomic introgressions from very distant lineages in a ciliate

Florian Bénitière (1), Olivier Arnaiz (2), Simon Penel (1), Sandra Duharcourt (3), Eric Meyer (4), Linda Sperling (2), Laurent Duret* (1)

(1) Universite Claude Bernard Lyon 1, LBBE, UMR 5558, CNRS, VAS, Villeurbanne, France (2) Université Paris-Saclay, CEA, CNRS - Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette, France (3) Université Paris Cité, CNRS, Institut Jacques Monod, F-75013 Paris, France (4) Institut de Biologie de l'ENS (IBENS), Département de Biologie, École Normale Supérieure, CNRS, INSERM, Université PSL, Paris, France

(*) corresponding author

Supplementary Tables

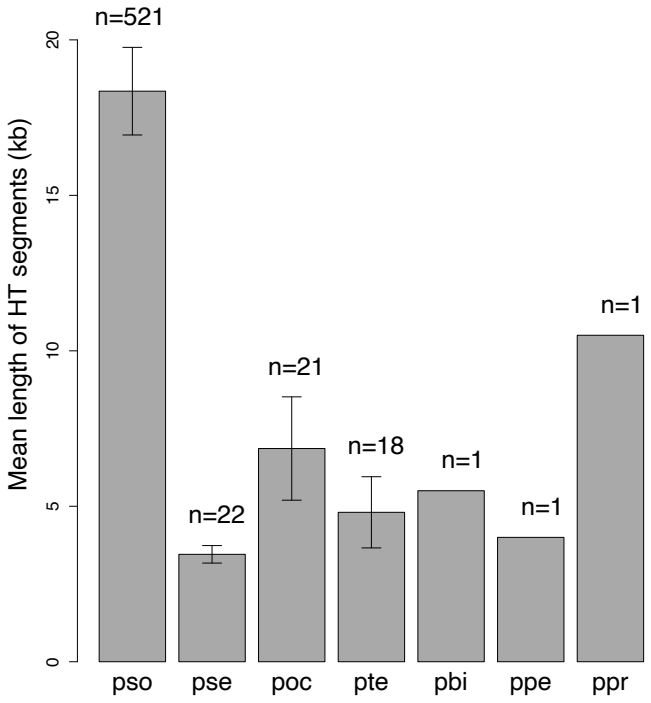
Donor species	HT event ID	Donor GeneID	Donor Gene Number	Donor scaffold	Psonn MIC contig	Psonn MAC scaffold	Location on Pson MAC scaffold	Psonn GeneID (MAC assembly)	Psonn Gene Number	Expression level in Psonn (vegetative cells, RPKM)	Comparison of Pson annotated CDS with its homologs	Functionality
Pquad	1	PQUADEC.NiA.1.G00030348	1	scaffold_0003	contig02903	scaffold_012	5'end	PSON.ATCC_30995.1.G0120001	1	21.132	Pson: 2 frameshifts: pseudogene	transcribed pseudogene
	2	PQUADEC.NiA.1.G00210133	2	scaffold_0021	contig06091	scaffold_098	middle	PSON.ATCC_30995.1.G0980057	2	1.770	Full-length CDS	Possibly functional
	2	PQUADEC.NiA.1.G00210134	3	scaffold_0021	contig06091	scaffold_098	middle	PSON.ATCC_30995.1.G0980055	3	1.129	Pquad homolog split into 2 genes in Psonn: pseudogene	transcribed pseudogene
	2	PQUADEC.NiA.1.G00210134	3	scaffold_0021	contig06091	scaffold_098	middle	PSON.ATCC_30995.1.G0980056	4	0.025	Pquad homolog split into 2 genes in Psonn: pseudogene	pseudogene
	2	PQUADEC.NiA.1.G00210134	3	scaffold_0021	contig06091	scaffold_098	middle	PSON.ATCC_30995.1.G0980053	5	0.030	Pson truncated + frameshift: pseudogene	pseudogene
	3	PQUADEC.NiA.1.G00950093	4	scaffold_0095	contig05439	scaffold_148	3'end	PSON.ATCC_30995.1.G1480144	6	0.038	Pson highly truncated: pseudogene	pseudogene
	3	PQUADEC.NiA.1.G00950094	5	scaffold_0095	contig05439	scaffold_148	3'end	PSON.ATCC_30995.1.G1480145	7	0.204	Full-length CDS	Intact CDS but very low exp.
	3	PQUADEC.NiA.1.G00950095	6	scaffold_0095	contig05439	scaffold_148	3'end	PSON.ATCC_30995.1.G1480146	8	3.104	Pson truncated/frameshifted: pseudogene	transcribed pseudogene
	3	PQUADEC.NiA.1.G00950097	7	scaffold_0095	contig06598	scaffold_148	3'end	PSON.ATCC_30995.1.G1480149	9	0.054	Pson truncated/frameshifted: pseudogene	pseudogene
	3	PQUADEC.NiA.1.G00950098	8	scaffold_0095	contig06598	scaffold_148	3'end	PSON.ATCC_30995.1.G1480150	10	2.493	Pson truncated/frameshifted: pseudogene	transcribed pseudogene
	3	PQUADEC.NiA.1.G00950098	8	scaffold_0095	contig06598	scaffold_148	3'end	PSON.ATCC_30995.1.G1480151	11	1.351	Pson truncated/frameshifted: pseudogene	transcribed pseudogene
	3	PQUADEC.NiA.1.G00950100	9	scaffold_0095	contig06306	scaffold_148	3'end	PSON.ATCC_30995.1.G1480155	12	0.046	Pson truncated/frameshifted: pseudogene	pseudogene
	3	PQUADEC.NiA.1.G00950120	10	scaffold_0095	contig01703	scaffold_106	3'end	PSON.ATCC_30995.1.G1060207	13	0.142	Full-length CDS	Intact CDS but very low exp.
	4	PQUADEC.NiA.1.G01090001	11	scaffold_0109	contig01786	scaffold_090	middle	PSON.ATCC_30995.1.G0900064	14	0.010	Full-length CDS	Intact CDS but very low exp.
	4	PQUADEC.NiA.1.G01090002	12	scaffold_0109	contig01786	scaffold_090	middle	PSON.ATCC_30995.1.G0900065	15	0.007	Full-length CDS	Intact CDS but very low exp.
	4	PQUADEC.NiA.1.G01090003	13	scaffold_0109	contig01786	scaffold_090	middle	PSON.ATCC_30995.1.G0900066	16	2.367	Pson highly truncated: pseudogene	transcribed pseudogene
	4	PQUADEC.NiA.1.G01090004	14	scaffold_0109	contig01786	scaffold_090	middle	PSON.ATCC_30995.1.G0900068	17	1.435	Pson truncated + frameshift :pseudogen	transcribed pseudogene
	4	PQUADEC.NiA.1.G01090004	14	scaffold_0109	contig01786	scaffold_090	middle	PSON.ATCC_30995.1.G0900067	18	0.937	Pson truncated + frameshift :pseudogen	transcribed pseudogene
Psex	5	PSEX.AZ8_4.1.G0110397	15	scaffold_011	contig03764	scaffold_049	middle	PSON.ATCC_30995.1.G0490038	19	0.409	Full-length CDS	Intact CDS but very low exp.
	6	PSEX.AZ8_4.1.G0440006	16	scaffold_044	contig00434	scaffold_036	5'end	PSON.ATCC_30995.1.G0360006	20	3.305	Psex homolog split into 2 genes/frameshifted: pseudogene	transcribed pseudogene
	6	PSEX.AZ8_4.1.G0440006	16	scaffold_044	contig00434	scaffold_036	5'end	PSON.ATCC_30995.1.G0360007	21	2.243	Psex homolog split into 2 genes/frameshifted: pseudogene	transcribed pseudogene
	6	PSEX.AZ8_4.1.G0440007	17	scaffold_044	contig00434	scaffold_036	5'end	PSON.ATCC_30995.1.G0360008	22	69.058	Full-length CDS	Possibly functional; high expr.
	7	PSEX.AZ8_4.1.G0930181	18	scaffold_093	contig03962	scaffold_043	middle	PSON.ATCC_30995.1.G0430234	23	0.246	Full-length CDS	Intact CDS but very low exp.
	7	PSEX.AZ8_4.1.G0930182	19	scaffold_093	contig03962	scaffold_043	middle	PSON.ATCC_30995.1.G0430232	24	0.829	Psex homolog split into 2 genes/frameshifted: pseudogene	transcribed pseudogene
	7	PSEX.AZ8_4.1.G0930182	19	scaffold_093	contig03962	scaffold_043	middle	PSON.ATCC_30995.1.G0430233	25	1.209	Psex homolog split into 2 genes/frameshifted: pseudogene	transcribed pseudogene
	7	PSEX.AZ8_4.1.G0930183	20	scaffold_093	contig03962	scaffold_043	middle	PSON.ATCC_30995.1.G0430231	26	0.961	Pson highly truncated: pseudogene	transcribed pseudogene
	8	PSEX.AZ8_4.1.G1000053	21	scaffold_100	contig01697	scaffold_034	3'end	PSON.ATCC_30995.1.G0340367	27	0.249	Psex gene split into 2 genes	pseudogene
	8	PSEX.AZ8_4.1.G1000053	21	scaffold_100	contig01697	scaffold_034	3'end	PSON.ATCC_30995.1.G0340368	28	0.042	Psex gene split into 2 genes	pseudogene
Ptred	9	PTRED.209.2.G71800001293250152	22	scf7180000129325	contig02752	scaffold_051	3'end	PSON.ATCC_30995.1.G0510351	29	1.144	Pson truncated/frameshifted: pseudogene	transcribed pseudogene
	9	PTRED.209.2.G71800001293250152	22	scf7180000129325	contig02752	scaffold_051	3'end	PSON.ATCC_30995.1.G0510350	30	1.056	Pson truncated/frameshifted: pseudogene	transcribed pseudogene
	9	PTRED.209.2.G71800001293250152	22	scf7180000129325	contig02752	scaffold_051	3'end	PSON.ATCC_30995.1.G0510352	31	0.736	Pson truncated/frameshifted: pseudogene	transcribed pseudogene
	9	NA	23	scf7180000129325	contig02752	scaffold_051	3'end	PSON.ATCC_30995.1.G0510347	32	0.137	No homolog annotated in Ptred	very low expression level
	9	NA	24	scf7180000129325	contig02752	scaffold_051	3'end	PSON.ATCC_30995.1.G0510346	33	0.054	No homolog annotated in Ptred	very low expression level
	9	PTRED.209.2.G71800001293250155	25	scf7180000129325	contig02752	scaffold_051	3'end	PSON.ATCC_30995.1.G0510345	34	0.807	Pson highly truncated: pseudogene	transcribed pseudogene
	9	PTRED.209.2.G71800001293250155	25	scf7180000129325	contig02752	scaffold_051	3'end	PSON.ATCC_30995.1.G0510349	35	0.196	Pson truncated/frameshifted: pseudogene	pseudogene
	9	PTRED.209.2.G71800001293250155	25	scf7180000129325	contig02752	scaffold_051	3'end	PSON.ATCC_30995.1.G0510348	36	0.065	Pson truncated/frameshifted: pseudogene	pseudogene
	9	PTRED.209.2.G71800001293250156	26	scf7180000129325	contig02752	scaffold_051	3'end	PSON.ATCC_30995.1.G0510344	37	1.432	Pson truncated/frameshifted: pseudogene	transcribed pseudogene
	9	PTRED.209.2.G71800001293250157	27	scf7180000129325	contig02752	scaffold_051	3'end	PSON.ATCC_30995.1.G0510343	38	0.001	Pson truncated/frameshifted: pseudogene	pseudogene
	9	PTRED.209.2.G71800001293250158	28	scf7180000129325	contig02752	scaffold_051	3'end	PSON.ATCC_30995.1.G0510342	39	0.009	Pson highly truncated: pseudogene	pseudogene
	10	PTRED.209.2.G71800001293360175	29	scf7180000129336	contig00726	scaffold_002	3'end	PSON.ATCC_30995.1.G0020587	40	2.758	Pson truncated/frameshifted: pseudogene	transcribed pseudogene
	10	PTRED.209.2.G71800001293360176	30	scf7180000129336	contig00726	scaffold_002	3'end	PSON.ATCC_30995.1.G0020589	41	3.943	Ptred homolog split into 3 genes in Psonn, with frameshifts -> pseudogene	transcribed pseudogene
	10	PTRED.209.2.G71800001293360176	30	scf7180000129336	contig00726	scaffold_002	3'end	PSON.ATCC_30995.1.G0020588	42	1.619	Ptred homolog split into 3 genes in Psonn, with frameshifts -> pseudogene	transcribed pseudogene
	10	PTRED.209.2.G71800001293360176	30	scf7180000129336	contig00726	scaffold_002	3'end	PSON.ATCC_30995.1.G0020590	43	1.613	Ptred homolog split into 3 genes in Psonn, with frameshifts -> pseudogene	transcribed pseudogene
	10	PTRED.209.2.G71800001293360177	31	scf7180000129336	contig00726	scaffold_002	3'end	PSON.ATCC_30995.1.G0020591	44	1.019	Full-length CDS	Possibly functional
	10	PTRED.209.2.G71800001293360178	32	scf7180000129336	contig00726	scaffold_002	3'end	PSON.ATCC_30995.1.G0020592	45	0.139	Pson highly truncated: pseudogene	pseudogene
	10	PTRED.209.2.G71800001293360179	33	scf7180000129336	contig04884	scaffold_002	3'end	PSON.ATCC_30995.1.G0020594	46	0.692	Pson truncated/frameshifted: pseudogene	transcribed pseudogene
	10	PTRED.209.2.G71800001293360182	34	scf7180000129336	contig04884	scaffold_002	3'end	PSON.ATCC_30995.1.G0020597	47	4.673	Full-length CDS	Possibly functional
	10	PTRED.209.2.G71800001293360183	35	scf7180000129336	contig04884	scaffold_002	3'end	PSON.ATCC_30995.1.G0020598	48	0.001	Pson truncated/frameshifted: pseudogene	pseudogene

Suppl. Table S1: **List of genes identified in HT segments located in the MAC-destined compartment of *P. sonneborni***

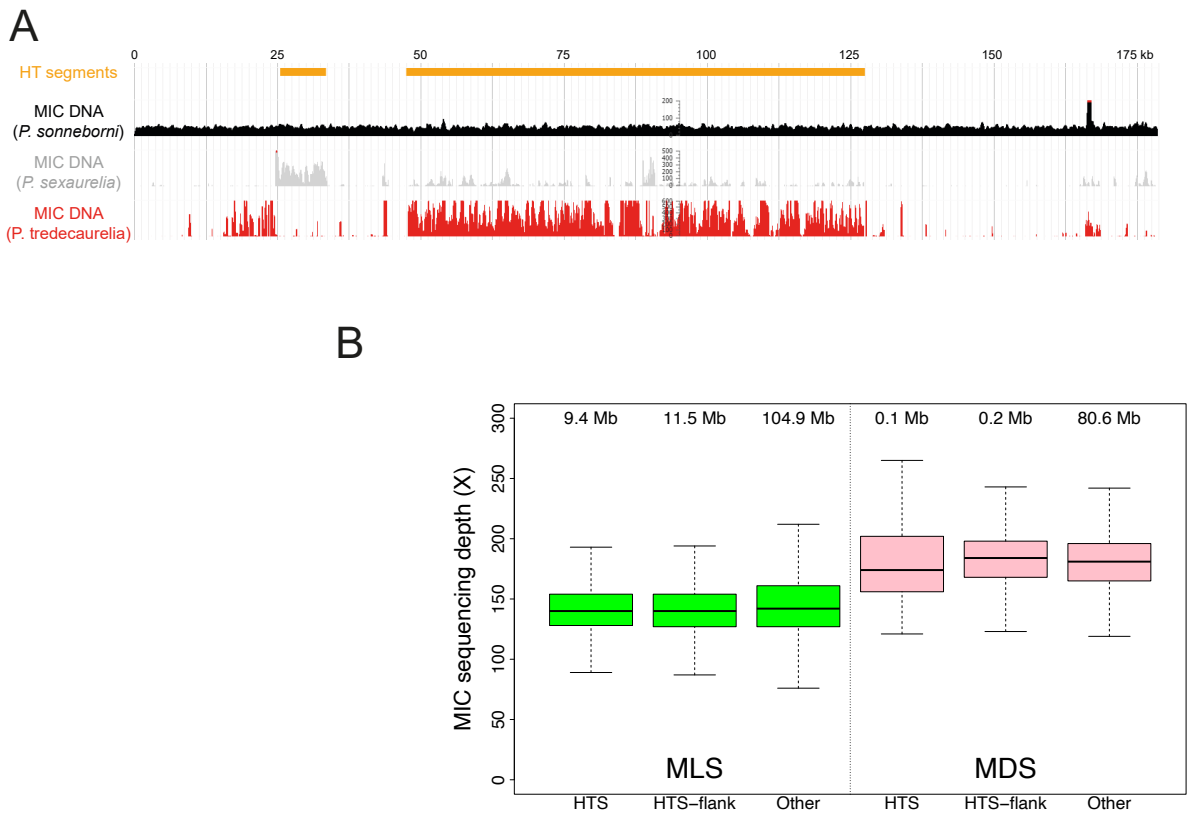
Species	Species code	Assembly name	Category	Strain	NCBI Assembly Accession/ParameciumDB link
P. biaurelia	*pbia*	pbiaurelia_mac_V1-4_v1.0	MAC	V1-4	https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/biaurelia/V1-4/sequences/
P. biaurelia	*pbia*	pbiaurelia_mic_V1-4_v1.0	MIC	V1-4	GCA_980646935.1
P. decaurelia	*pdec*	pdecaurelia_mac_223_v1.0	MAC	223	https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/decaurelia/223/sequences/
P. dodecaurelia	*pdod*	pdodecaurelia_mac_274_v1.0	MAC	274	https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/dodecaurelia/274/sequences/
P. jenningsi	*pjen*	pjenningsi_mac_M_v1.0	MAC	M	https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/jenningsi/M/sequences/
P. novaurelia	*pnov*	pnovaurelia_mac_TE_v1.0	MAC	TE	https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/novaurelia/TE/sequences/
P. octaurelia	*poct*	poctaurelia_mac_138_v1.0	MAC	138	https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/octaurelia/138/sequences/
P. octaurelia	*poct*	poctaurelia_mic_138_v1.1	MIC	138	GCA_980644265.1
P. pentaurelia	*ppen*	ppentaurelia_mac_87_v1.0.MAC-constitutive	MAC	87	https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/pentaurelia/87/sequences/
P. pentaurelia	*ppen*	ppentaurelia_mic_87_v1.1	MIC	87	GCA_980644265.1
P. primaurelia	*ppri*	pprimaurelia_mac_AZ9-3_v1.1.MAC-constitutive	MAC	AZ9-3	https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/primaurelia/AZ9-3/sequences/
P. primaurelia	*ppri*	pprimaurelia_mic_AZ9-3_v1.0	MIC	AZ9-3	GCA_980642815.1
P. quadecaurelia	*pqua*	pquadecaurelia_mac_NiA_v1.0	MAC	NiA	https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/quadecaurelia/NiA/sequences/
P. sexaurelia	*psex*	psexaurelia_mac_AZ8-4_v1.1	MAC	AZ8-4	https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/sexaurelia/AZ8-4/sequences/
P. sexaurelia	*psex*	psexaurelia_mic_AZ8-4_v1.1	MIC	AZ8-4	GCA_980645855.1
P. sonneborni	*pson*	psonneborni_mac_ATCC_30995_v1.0.MAC-constitutive	MAC	ATCC_30995	https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/sonneborni/ATCC_30995/sequences/
P. sonneborni	*pson*	psonneborni_mic_ATCC_30995_454Contigs	MIC	ATCC_30995	GCA_980648905.1
P. tetraurelia	*ptet*	ptetraurelia_mac_51_v2.0.MAC-constitutive	MAC	51	https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/tetraurelia/51/sequences/
P. tetraurelia	*ptet*	ptetraurelia_mic_51_v1.2	MIC	51	https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/tetraurelia/51/sequences/
P. tredecaurelia	*ptre*	ptredecaurelia_209_AP38_filtered	MAC	209	https://paramecium.i2bc.paris-saclay.fr/download/Paramecium/tredecaurelia/209_AP38/sequences/

Suppl. Table S2: **List of genome assemblies and strains used in this study**

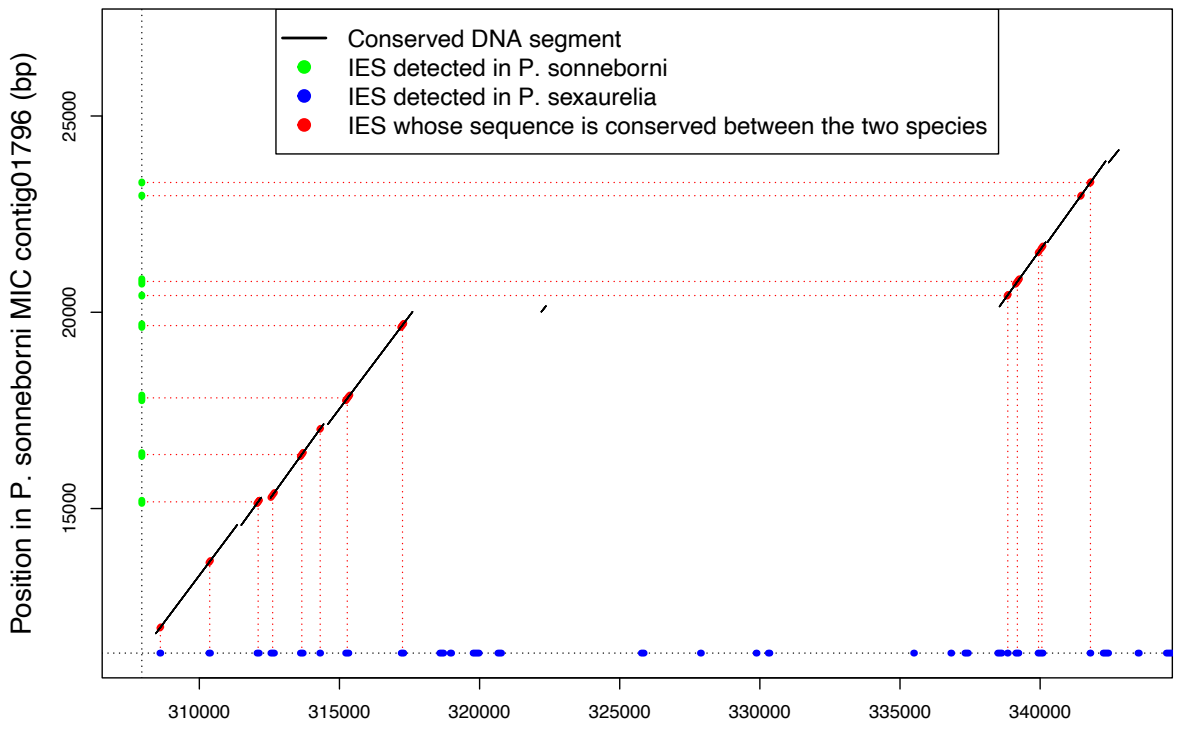
Supplementary Figures



Suppl. Figure S1: **Average length of HT segments detected in the MIC genomes of the 7 analyzed species**. Error bars represent the 95% confidence interval for the mean. Species codes: pso: *P. sonneborni*, pte: *P. tetraurelia*, pbi: *P. biaurelia*, poc: *P. octaurelia*, pse: *P. sexaurelia*, ppr: *P. primaurelia*, ppe: *P. pentaurelia*.



Suppl. Figure S2: **Evidence that HT segments detected in *P. sonneborni* do not result from DNA contamination artefacts**. **A.** Example of a MIC contig of *P. sonneborni* (contig00015) that includes two HT segments (orange rectangles), the first derived from *P. sexaurelia* (scaffold_001.2), the second from *P. tredecaurelia* (scf7180000129386). If HT segments resulted from DNA contamination, this would imply that contig00015 corresponds to an artefactual chimeric assembly of DNA fragments coming from three sources (*P. sonneborni*, *P. sexaurelia* and *P. tredecaurelia*). Hence, the sequencing depths would be expected to differ between the two HT segments (originating from *P. sexaurelia* and *P. tredecaurelia*), and to be lower in HT segments than in flanking regions (*P. sonneborni*). The first profile (black) indicates the sequencing depth for the *P. sonneborni* MIC DNA sample. This profile shows a uniform sequencing depth along the whole contig, including in regions outside of HT segments, in contradiction with the hypothesis of a DNA contamination artefact. As a control, we also mapped sequence reads from MIC DNA samples obtained from *P. sexaurelia* (grey) or *P. tredecaurelia* (red). As expected, we observed a high density of reads from each species mapping on the corresponding HT segment. The heterogeneity of coverage reflects the fact that the sequences of the *P. sexaurelia* and *P. tredecaurelia* strains that we used in this study are not identical to the HT segments present in *P. sonneborni*, again in contradiction with the hypothesis that these HT segments might result from DNA contamination. **B.** Distribution of sequencing depth within and outside HT segments in *P. sonneborni*. The MIC genome assembly was split into 500bp-long non-overlapping windows, which were classified as MIC-limited (MLS) or MAC-destined (MDS), based on their MAC sequencing depth, and then subdivided into three subsets: windows that overlap a HT-segment (HTS), windows that flank a HT-segment (HTS-flank), and windows located on other MIC contigs (i.e. not flanking a HT-segment). Both in MLS and in MDS regions, the distribution MIC sequencing depth (represented by boxplots) is the same within HT-segments as in their flanking regions and the same as in the rest of the genome, consistent with the assumption that HT-segments correspond to authentic *P. sonneborni* MIC DNA. The amount of DNA in each class of windows is given above the corresponding boxplot.



Suppl. Figure S3: **Presence of conserved IESs within HT segments**. This figure presents a dot plot alignment between a *P. sonneborni* MIC contig (contig01796) and the *P. sexaurelia* MIC scaffold (scaffold_103), with which we identified a HT segment (88% sequence identity between the two species). IESs have been previously detected in the MIC genomes of *P. sonneborni* and of *P. sexaurelia* (Sellis et al. 2021). The figure shows that this HT segment encompasses 14 conserved IESs, among which 7 were detected in both species, 6 were detected only in *P. sexaurelia* and 1 only in *P. sonneborni*. Given that IESs are absent from the MAC genome, the presence of conserved IESs within the HT segment indicates that it resulted from the transfer of MIC, rather than MAC DNA.