SCIENCE CHINA PRESS

SCIENCE CHINA Life Sciences
life.scichina.com link.springer.com
https://doi.org/10.1007/s11427-024-2872-7

RESEARCH PAPER

Whole-genome duplications revealed by macronuclear genomes of five rare species of the model ciliates *Paramecium*

Jiahao Ni[1,2], Yue Hao[3,4], Berenice Jiménez-Marín[3], Farhan Ali[3], Jiao Pan[1], Yaohai Wang[1], Ziguang Deng[1], Jean-Francois Gout[5], Yu Zhang[1], Michael Lynch[3*] & Hongan Long[1,2*]

[1]Key Laboratory of Evolution and Marine Biodiversity (Ministry of Education), Institute of Evolution and Marine Biodiversity, Ocean University of China, Qingdao 266003, China
[2]Laboratory for Marine Biology and Biotechnology, Qingdao Marine Science and Technology Center, Qingdao 266237, China
[3]Biodesign Center for Mechanisms of Evolution, Arizona State University, Tempe 85287, USA
[4]Cancer and Cell Biology Division, Translational Genomics Research Institute, Phoenix 85004, USA
[5]Department of Biological Sciences, Mississippi State University, Mississippi 39762, USA

*Corresponding authors (Hongan Long, email: longhongan@ouc.edu.cn; Michael Lynch, email: mlynch11@asu.edu)

Received 1 November 2024; Accepted 21 January 2025; Published online 15 August 2025

***Paramecium*, a group of ciliates with a long evolutionary history, plays essential roles in freshwater ecosystems and has been model for genetic, cellular, and evolutionary studies for over a century. Despite the valuable contributions of genomic resources such as ParameciumDB, genomic data are still mostly limited to species in and near the *P. aurelia* group. This study addresses this gap by HiFi sequencing, assembling, and annotating the macronuclear genomes of five rare *Paramecium* species: *P. calkinsi*, *P. duboscqui*, *P. nephridiatum*, *P. putrinum*, and *P. woodruffi*. These genomes enable a comprehensive exploration of genomic diversity, genome evolution, and phylogenomic relationships within the genus *Paramecium*. The genome sizes range from 47.78 to 113.16 Mb, reflecting unexpected variation in genomic content, and genic features differ from those of other reported *Paramecium* genomes, such as larger intron sizes and higher GC content. Nonetheless, the *de novo* assemblies indicate that macronuclear genomes of all *Paramecium* are highly streamlined, with ~77% being protein-coding gene regions. Based on gene-duplication depths, synonymous mutations in paralogs, and phylogenomic relationships, we discovered that the five species experienced at least three whole-genome duplication (WGD) events, independent of those previously found in the *P. aurelia* complex. Using all available WGD data for *Paramecium*, we further explore the paralog dynamics after WGD events by modeling. This study contributed to a more comprehensive and deeper understanding of genome architecture and evolution in *Paramecium*.**

ciliate | genome diversity | macronucleus | *Paramecium* | protist

INTRODUCTION

Rare species, despite their limited numbers and distribution, can have a significant impact on ecological communities by maintaining diversity and vital ecological functions. For example, in coral reefs, alpine regions, and tropical forests, rare species often support essential ecological roles that more abundant species do not fulfill (Mouillot et al., 2013). Among ciliates, many rare species have been reported, including those in the genus *Paramecium*. *Paramecia* are model organisms well-known for their contributions to understanding mating type, cytoplasmic inheritance, genome rearrangement, whole-genome duplication (WGD) so on (Aury et al., 2006; Beisson and Sonneborn, 1965; Garnier et al., 2004; Gout et al., 2023; Johri et al., 2022; Kimball, 1943; Sonneborn, 1937). Most of what we know about *Paramecium* is limited to the *P. aurelia* species complex (Johri et al., 2017; Long et al., 2023; Samuel et al., 1981), whereas *P. woodruffi* and most other non-*aurelia Paramecium* have only been documented through morphological observation and marker genes, with little genome-level research (Fokin, 2010; Long et al., 2023; Sonneborn, 1970; Wichterman, 2012). The genomic resources of the latter species will help test or complement what we discovered in the former, contributing to a comprehensive and robust understanding of *Paramecium* evolution at the genome level.

The first macronuclear genome of *Paramecium* (*P. tetraurelia*), published in 2006, revealed three rounds of WGD in the phylogenetic history of this model organism, with an extremely long evolutionary history dating back 500 Mya (Aury et al., 2006). WGDs have been reported in many other sequenced organisms, such as hexapods, the allotetraploid frog *Xenopus laevis*, and rainbow trout (Berthelot et al., 2014; Li et al., 2018; Session et al., 2016). However, it was not until 2023 that the WGDs and gene-retention patterns of other members of the *P. aurelia* group were revealed (Gout et al., 2023). According to the ParameciumDB (as of March 2024), 22 macronuclear genomes of 15 *Paramecium* species have been published. However, the majority of genomic resources are of the *P. aurelia* species

Citation: Ni, J., Hao, Y., Jiménez-marín, B., Ali, F., Pan, J., Wang, Y., Deng, Z., Gout, J.F., Zhang, Y., Lynch, M., et al. (2025). Whole-genome duplications revealed by macronuclear genomes of five rare species of the model ciliates *Paramecium*. Sci China Life Sci 68, 3633–3645. https://doi.org/10.1007/s11427-024-2872-7

complex, and a few other easily-collected species, such as *P. bursaria* containing green *Chlorella* symbionts, and *P. caudatum* and *P. multimicronucleatum* with large cell size (Cheng et al., 2020; Gout et al., 2023; McGrath et al., 2014). Studies on population genetics and *Paramecium* diversity are also limited to the above few *Paramecium* species (Catania et al., 2008; Johri et al., 2017), and most *Paramecium* species still lack genome and transcriptome information (Potekhin and Mayén-Estrada, 2020).

Despite their utility as eukaryotic model organisms, the phylogenetic relationships of *Paramecium* species have not yet been fully resolved. Questions regarding the basal species in *Paramecium* phylogenetic evolution, which species are included in the *P. woodruffi* subgenus, and the paralog dynamics after WGDs, remain open (Boscaro et al., 2012; Fokin and Chivilev, 2000; Fokin et al., 2001; Greczek-Stachura et al., 2021; Strüder-Kypke et al., 2000; Woodruff, 1921). Apart from the *P. aurelia* sibling species complex, most phylogenetic trees are constructed using 1–3 gene markers, with limited numbers of informative sites (Boscaro et al., 2012; Gout et al., 2023; Long et al., 2023; Schrallhammer et al., 2006; Strüder-Kypke et al., 2000; Tarcz et al., 2014). The limitations of using morphogenesis and single or multiple molecular markers for determining phylogenetic relationships are evident. Phylogenomic analyses, using high-quality genomes of diverse species, thus offer a promising methodology to address these challenges.

In terms of genome diversity and evolutionary patterns, recent studies have demonstrated high diversity in gene family expansion, genome fragmentation, and regulatory mechanisms of ciliates (Catania et al., 2021; Cheng et al., 2020; Gout et al., 2023; Jin et al., 2023; Johri et al., 2022; Li et al., 2021; Lyu et al., 2024; Sellis et al., 2021; Yan et al., 2019; Zhang et al., 2023; Zheng et al., 2020; Zheng et al., 2021). These findings lay the foundation for our understanding of *Paramecium* genome evolution. However, genomic studies on rare *Paramecium* species remain relatively scarce. The full picture of WGDs in *Paramecium*, the mechanisms of species differentiation, genome divergence, and the key factors determining genome size and the diversity and evolution of genome architecture, all require more genome resources, especially those of rare species.

There are many previously described but understudied *Paramecium* species, such as *P. calkinsi*, *P. duboscqui*, *P. nephridiatum*, *P. putrinum*, and *P. woodruffi* (Fokin et al., 1999; Przyboś et al., 2019; Sabaneyeva, 1997), for all of which genetic and genomic information is mostly absent. All of them only have a few collection records in the literature, and molecular studies of multiple strains of *P. putrinum* were not even available until 2014 (Fokin, 2010; Fokin et al., 2001; Tarcz et al., 2014). Among these species, three of them possibly belong to the *P. woodruffi* group, which includes *P. nephridiatum*, *P. woodruffi*, and *P. calkinsi* (Fokin and Chivilev, 2000; Woodruff, 1921). *P. putrinum* is not a member of the *P. woodruffi* group, which is also supported by its multiple mating-type systems (Fokin et al., 2001; Jankowski, 1972), and neither is *P. duboscqui*, reflected by its distinct nuclear reorganization processes (Fokin et al., 2001). There are also reports of cryptic species (genetically but not morphologically distinguishable) in both *P. putrinum* and *P. duboscqui* (Boscaro et al., 2012; Schrallhammer et al., 2006; Tarcz et al., 2014), but these studies have consistently lacked power in molecular evidence for phylogenetic inference and cryptic species identification. Much of the confusion could be addressed with whole genome information.

Isolates of five rare *Paramecium* species (*P. calkinsi* Woodruff 1921 GN5-3, *P. duboscqui* Chatton and Brachon 1933 PD1, *P. nephridiatum* Gelei 1925 Rw-1, *P. putrinum* Claparéde and Lachman 1858 OM4 and *P. woodruffi* Wenrich 1928 GNS4), which were maintained at Yamaguchi University under the National BioResource Project, were used in this study. We performed macronuclear isolation and PacBio HiFi sequencing and assembled and annotated their high-quality macronuclear genomes. Together with previously published genome data, we explore the genomic diversity, genome evolution, phylogenomic relationships, and paralog dynamics after WGD of *Paramecium*.

RESULTS

Highly streamlined macronuclear genomes of the five rare *Paramecium* species

We first isolated the macronuclei of the five species, extracted macronuclear DNA and total RNA, and applied PacBio HiFi sequencing, with a coverage depth of 64× to 127× after HiFi reads conversion (Table 1). We *de novo* assembled the genomes using Canu after multiple filtering steps (see MATERIALS AND METHODS). The genome sizes of these *Paramecium* species vary from 47.78 to 113.16 Mb. Among them, *P. putrinum* has the largest genome among all studied *Paramecium* species, nearly four times larger than that of *P. bursaria* (26.82 Mb, the smallest genome in *Paramecium*) (Cheng et al., 2020).

The mapping rates of the HiFi raw reads to the assemblies all exceeded 95% (Table 1), indicating the high purity of the genomic DNA of the isolated macronuclei. As with most previously published macronuclear genomes of *Paramecium* (Arnaiz et al., 2020), scaffolds with both telomeres in the five assemblies are rare, but 17%–66% of them contain at least one telomere, indicating that a large fraction of the scaffolds cover entire chromosome arms (Table S1). We further examined the proportion of HiFi reads containing telomere sequences, which is far fewer than that of telomere-containing scaffolds in the genome (Table S1). This demonstrated that the lack of scaffolds with two telomeres was possibly caused by the relatively short HiFi reads obtained from sequencing these extremely AT-rich genomes (Table 1), not assembling or filtering.

Taking advantage of the published *Paramecium* annotation data and RNAseq data generated in this study, we successfully constructed a weighted array model (WAM) and annotated the five genomes using EuGene (Sallet et al., 2019). Only 29 exons skipping events and 3 exons mutually exclusive events were found in *P. nephridiatum*, accounting for 0.09% of the genes in this species. No significant alternative splicing signs were observed in the other four species. The gene numbers in the five species range from 21,026 to 50,109 (Table 2; Table S2), with the highest number of genes found in *P. putrinum*, consistent with its largest genome size. There is significant variation in the distributions of gene length ($P<0.008$, Kolmogorov-Smirnov test after Bonferroni correction) (Figure 1). Compared with *P. tetraurelia* with three WGDs, these five species have a greater number of shorter chromosomes and exhibit richer collinearity within their genomes, suggesting the possible occurrence of WGDs (Figure 2). Such WGDs can lead to the expansion and subsequent diversification of gene families. To further understand the genetic relationships and the impact of these duplications, we explored gene families and created Venn

Table 1. Assembly statistics of the five *de novo* assemblies

Species	Genome size (bp)	Mapping rate (%)	GC content (%)	Contig number	N50	L50	Median length of HiFi reads (bp)
P. calkinsi GN5-3	62,445,650	97.16	27.87	1,460	47,291	295	13,285
P. duboscqui PD1	47,782,392	98.29	32.02	770	172,616	51	13,860
P. nephridiatum Rw-1	89,038,773	95.61	30.75	2,221	43,929	404	15,854
P. putrinum OM4	113,155,677	99.46	32.91	3,071	40,706	872	13,450
P. woodruffi GNS4	70,726,321	97.08	29.16	1,583	40,994	214	13,630

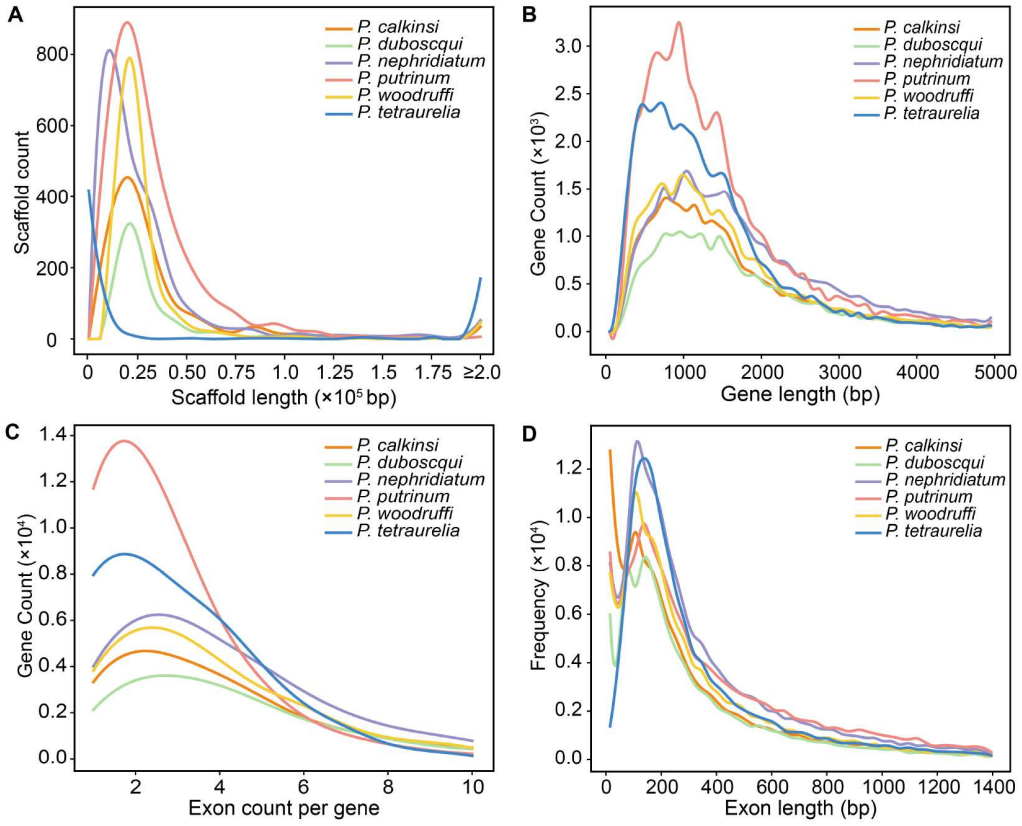

Figure 1. Statistics of genome structure in the five rare *Paramecium* and *P. tetraurelia*. A, Scaffold length distribution, with a bin size of 10,000 bp, scaffolds longer than 200,000 bp are merged. B, Gene length distribution with a bin size of 100 bp, genes longer than 5 kb are extremely scarce and thus not depicted. The distribution across species significantly differs (Kolmogorov-Smirnov test with Bonferroni correction, $P \leqslant 0.008$) in terms of gene length. C, Distribution of exon number per gene, excluding genes with more than 10 exons; the curves have been smoothed. D, Exon length distribution, with a bin size of 30 bp.

diagrams for those shared among the five rare species, and identified unique ones for all *Paramecium* species (Figure 3). Of the five rare species, *P. putrinum* has 10,000 to 20,000 more genes than the other four species, and most of these genes are in the 2,525 non-homologous gene families (12,051 genes). The biological functions of these genes are shown in Table S3. The *de novo* assemblies from these rare *Paramecium* species have increased the number of species-specific gene families within the genus to 5,473, a significant rise from the 1,607 unique gene families if only considering previously published *Paramecium* genomes.

Notably, *P. nephridiatum* has the longest mean gene length of ~2,045 bp, which is about 616 bp longer than that of *P. tetraurelia* (Tables S2 and S4). Yet, the proportions of total coding gene length in the genomes of these two species are close, i.e., 80.78% in *P. nephridiatum* and 78.58%–79.99% in *P. tetraurelia* (Aury et al., 2006; Sellis et al., 2021). Examination of the macronuclear genome in other *Paramecium* species revealed a similar total coding gene length proportion around 77%, with the highest of ~88.94% in *P. caudatum* (Figure S1, Tables S2 and S4). Considering the divergence of the *Paramecium* genus possibly dating back to 579.5–1,256.4 Mya (Kumar et al., 2017; Parfrey et al., 2011; Rataj and Vďačný, 2018), it appears that *Paramecium* macronuclear genomes have been persistently highly streamlined, with short intergenic regions and high gene density, in strong contrast to most other eukaryotic genomes.

The introns of the five rare *Paramecium* species are significantly longer than those of previously reported *Paramecium* genomes (mean size range: 31.48–42.43 bp *vs.* 25.64–26.66 bp; *t*-test, P=0.0051), which are still much shorter than *Arabidopsis* (~168 bp), human (~6,398 bp), and *Saccharomyces cerevisiae* (100–400 bp) (Tables S2 and S4) (Chang et al., 2017; Piovesan et al., 2019;

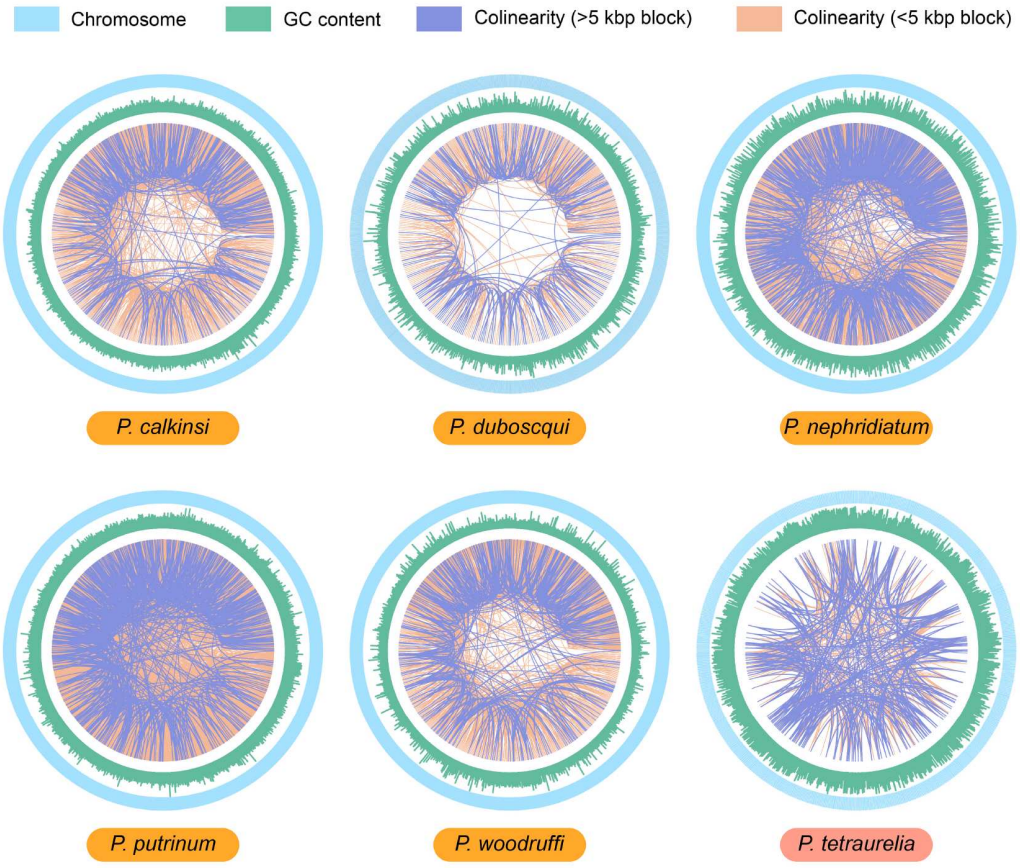

Figure 2. Genomic structure comparisons between five rare *Paramecium* species and *P. tetraurelia*. From the outer to the inner layers, the circles display chromosomes, GC content, and collinearity relationships within each genome. Collinear blocks longer than 5 kb are highlighted in blue (shorter ones in orange).

Spingola et al., 1999). As shown in Figure 4, the majority of the introns in the five newly assembled genomes range from 20 to 30 bp in length. In general, the intron length of most genes is less than 40% of the total gene length, although some genes have introns that occupy a high proportion of the gene length (Figure S2). Notably, the maximum number of introns per gene in *P. putrinum* and *P. woodruffi* is exceptionally low, whereas, in the genomes of the other three species, there are a few genes with over 50 introns (Figure S2), with their functions being unclear.

Consistent with previous findings on *P. tetraurelia*, most introns of the five new assemblies are not divisible by three (Figure 4) (Bondarenko et al., 2016). Only a small fraction of those divisible by three contain the TGA stop codon. Stopless 3*n* introns are in counter-selection because they cannot give rise to a premature termination codon (PTC) if retained in the transcript, so they are potentially deleterious. This consistency with prior research on *P. tetraurelia* underlines the importance of intron length and composition in the regulation of gene expression and the maintenance of protein integrity within the *Paramecium* genus. These findings contribute to our understanding of genomic organization and evolution, highlighting the variability and adaptive significance of intron characteristics among closely related species.

De novo assembled macronuclear genomes of the five rare species also reveal unforeseen variation in the genomic GC content of *Paramecium*. The distribution of the GC content of the scaffolds in each of the five genomes is highly different from that of the existing genomes, especially *P. duboscqui*, *P. nephridiatum*, and *P. putrinum*, whose mean GC% is higher than 30% (Figure S3, Tables S2 and S4). The GC% of the introns is lower than that of the overall genome in all *Paramecium* (Tables S2 and S4). When the GT-AG sequences at the ends of each intron are disregarded, the GC content of the introns further decreases, with some introns completely lacking guanines or cytosines (Figure 4). Similarly, high AT content also appears in exon regions, with a lower frequency of usage for codons with higher GC content (Figure S4). This is consistent with the high mutation bias in the AT direction of *Paramecium* (Long et al., 2018).

Phylogenomic relationships between *Paramecium* species and widespread WGDs

The phylogenetic evolution of *Paramecium* remains unclear, primarily due to the scarcity of whole-genome data. To address this, we constructed a phylogenomic tree using protein sequences from *de novo* assembled and previously published macronuclear genomes (Figure 5). Genes with incomplete reading frames or multiple internal stop codons were filtered out. The phylogenomic tree supports the monophyly and clustering of *P. aurelia* species (Figure 5). *P. calkinsi*, *P. nephridiatum*, and *P. woodruffi* cluster together, validating previous predictions of them being in the *P. woodruffi* group based on morphology and the nuclear reorganization process (Fokin and Chivilev, 2000; Fokin et al., 2001). *P. duboscqui*

https://doi.org/10.1007/s11427-024-2872-7

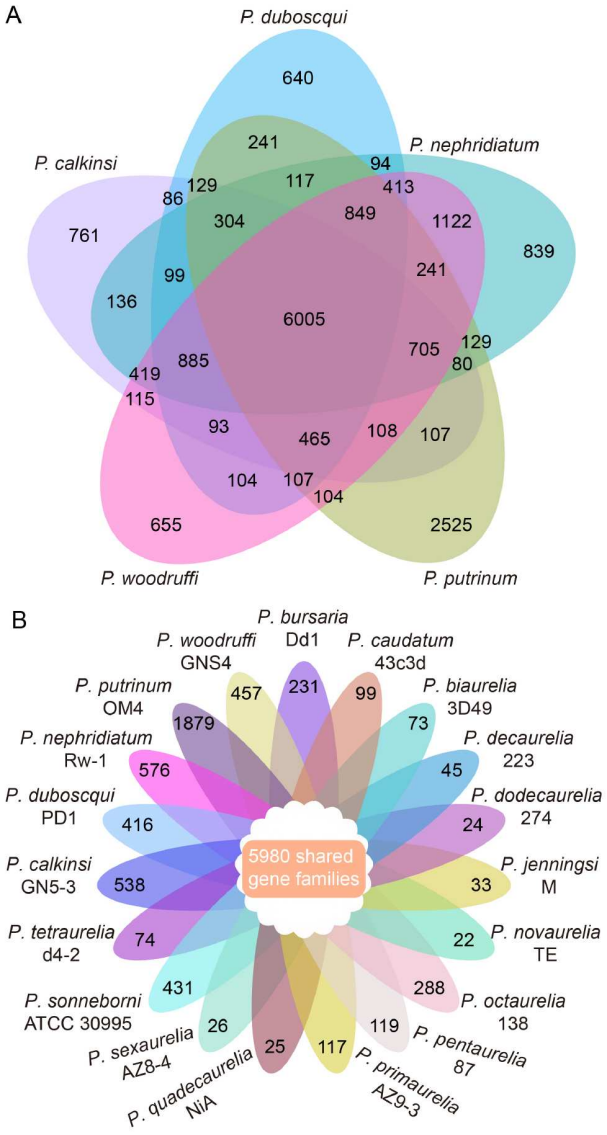

Figure 3. Gene family statistics. A, The shared gene families among the five rare species. B, The number of unique gene families for each *Paramecium* species.

shows obvious non-clustering with *P. woodruffi*, supporting Fokin's suspicion about their partially mismatched morphological features (Fokin and Chivilev, 2000; Fokin et al., 2001). Notably, like previous results from morphology and the 18S rDNA gene sequence (Boscaro et al., 2012; Strüder-Kypke et al., 2000), our study further demonstrates that *P. bursaria*, *P. duboscqui*, and *P. putrinum* are all potential basal species of *Paramecium* (Figure S5).

Both phylogenetic trees and previous morphological studies have shown that *P. calkinsi*, *P. nephridiatum*, and *P. woodruffi* belong to the *P. woodruffi* group (Fokin, 2010; Fokin et al., 2001), but their branch lengths differ greatly in Figure 5. We further determined whether the common ancestor of *P. nephridiatum* and *P. woodruffi* went through any WGD events, which did not occur in the *P. calkinsi* lineage, similar to the WGD scenarios of *P. caudatum* and *P. aurelia*. We conducted a statistical analysis of the gene families from the three species, identifying a total of 7,494 gene families containing genes from all species. Among these, 4,949 gene families had at least one species with multiple paralogs. Notably, in 4,452 of these 4,949 families (89.96%), genes from the same species were clustered into monophyletic groups, and the phylogenetic relationships within these families were consistent with the species tree. This analysis underscores the occurrence and sharing of WGD events. Collinearity analysis among the three species also supported such a WGD event, by showing signs of similar paralog number and number of annotated genes in both *P. nephridiatum* and *P. woodruffi* (Figure 6). *P. nephridiatum* and *P. woodruffi* share 39,004 collinear relationships, with 9,975 pairs showing the same paralog number within their genomes. By contrast, *P. calkinsi* has only 14,221 and 15,932 collinear relationships with *P. nephridiatum* and *P. woodruffi*, respectively, with about 4,000 shared genes having the same paralog number in each genome: approximately half that of *P. nephridiatum* with *P. woodruffi*. All of the above suggest that *P. nephridiatum* and *P. woodruffi* may share more WGD events, which appear to be absent in *P. calkinsi*. These three species share a more ancient WGD event (Figure 6). However, the branch length of *P. woodruffi* and *P. calkinsi*, calculated from their ancient WGD, almost doubles that of *P. caudatum* and *P. aurelia*, leaving fewer WGD remnants than in *P. aurelia*.

To determine the occurrence of WGD in the other two species, we analyzed gene families shared by all five species, totaling approximately 8,000 families. Among these, 5,981 gene families had at least one species with multiple genes. Notably, 5,083 of these families (84.99%) formed monophyletic groups, and the topologies of these groups were consistent with the species tree. Combined with collinearity figures, there are also signs of WGD in *P. duboscqui* and *P. putrinum*, further confirming that WGDs in the *Paramecium* genus are present across most of its subgenera (but with no evident WGD signals in *P. bursaria*).

To further confirm the occurrence of WGDs, we sought additional evidence to validate the occurrence of WGD events. Extensive genome duplication events result in the generation of numerous paralogous genes, which accumulate mutations separately. *Ks* (synonymous substitutions per site) of paralogous gene pairs can help detect WGDs (Lynch and Conery, 2000). Through self-blast alignment analysis using the longest collinear blocks, we calculated the *Ks* values for genes within the blocks (Figure S6). Multiple peaks confirm multiple WGDs. WGDs in

Table 2. Genome annotation statistics of the species sequenced in this study. Previously sequenced genomes and other detailed information are in Tables S2 and S4

Species	Gene number	Mean gene length (bp)	Total coding gene proportion (%)	Mean exon length (bp)	Mean intron length (bp)	Mean intron number	Total repeat sequence length (bp)
P. calkinsi GN5-3	26,043	1,749	71.31	355	33	3.59	6,911,006
P. duboscqui PD1	21,026	1,890	82.25	373	34	3.73	3,223,650
P. nephridiatum Rw-1	35,633	2,045	80.78	419	44	3.51	12,592,531
P. putrinum OM4	50,109	1,547	67.65	501	38	1.94	21,152,340
P. woodruffi GNS4	29,946	1,712	70.81	379	36	3.22	7,069,176

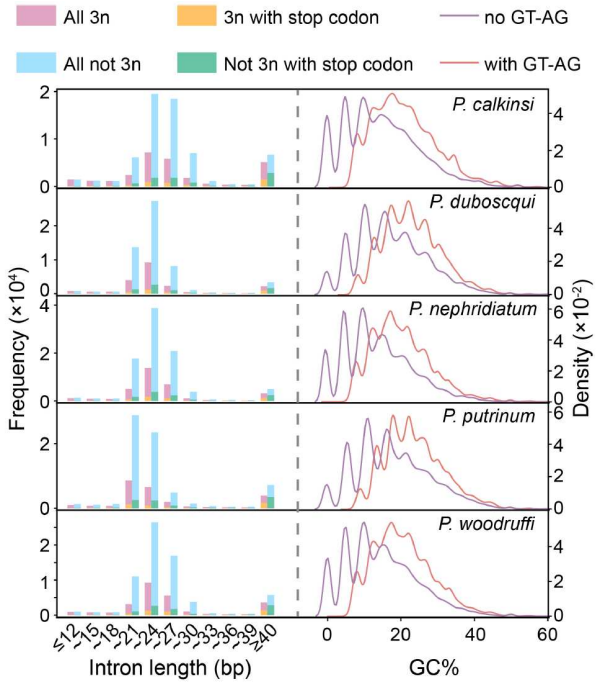

Figure 4. Intron length and GC content distributions of the five *de novo* assemblies. Left: purple and blue bars show the distributions of 3*n* and 3*n*±1 introns, respectively. Yellow and green bars show the distributions of 3*n* and 3*n*±1 introns with in-frame stop codons, respectively. Right: GC content density of introns (red) and GC content of introns without terminal GT/AG (purple).

Paramecium are thus more widespread than previously thought. Divergence in gene number and content after WGDs is thus possibly associated with genomic diversification or even speciation within the *Paramecium* genus.

Gene duplication depth shows evolutionary dynamics in *Paramecium*

If one gene within a collinear block (containing more than 5 genes) displays collinearity with *n* other genes (belonging to *n* different blocks), we consider the gene duplication depth of this gene to be *n* (simply understood as the number of paralogs for a gene). The maximum gene duplication depth of all genes indicates the total number of WGDs, since the overall gene count doubles after each WGD event, e.g., max=7 (8 if counting the original copy, 2^3=8), possibly represents three WGD events. Based on the Reciprocal Best Hit (RBH) sliding window method and ancestral genome inference, the *P. aurelia* species complex has been confirmed to have undergone three WGD events (Aury et al., 2006; Gout et al., 2023). The majority of *P. aurelia* lineages show a maximum duplication depth of 7 or 8, corresponding to the theoretical depth limit of three consecutive WGD events (2^3) (Figure 7). In *P. calkinsi*, *P. duboscqui*, *P. nephridiatum*, *P. putrinum*, and *P. woodruffi*, 64.58%, 66.10%, 74.46%, 73.53%, and 77.93% of genes were found to have signs of gene duplication. The maximum duplication depths in the genomes of the five rare species are generally higher than those of *P. aurelia*, implying comparable or more WGD events. However, certain species like *P. nephridiatum* show an exceptionally high maximum duplication depth, while the number of genes with extensive duplication is relatively small (Figure 7). The functions of most of the genes are still unclear, while others are mainly involved in biological processes such as signal transduction, metabolism, DNA repair, and immune response (Table S5). It is challenging to directly infer whether these duplicated genes were WGD remnants or from other small-scale duplication events.

After we removed the depths with <100 genes when determining the number of WGD events (by max duplication depth), all five rare *Paramecium* showed three WGD events (Figure 7). Species of *P. aurelia* and non-*P. aurelia*, which has undergone three rounds of WGDs, displays different distributions of gene duplication depths and is also confirmed to have consistent WGD numbers with those reported in previous studies (Aury et al., 2006; Gout et al., 2023; McGrath et al., 2014). In all species (except *P. caudatum*), there is a general trend of fewer genes at higher duplication depths (Figure 7). The branches of the five non-*P. aurelia* species, derived from the respective nodes before the most ancient WGD occurred, are generally longer, suggesting that extended evolutionary time may promote a convergence toward patterns of fewer genes at higher duplication

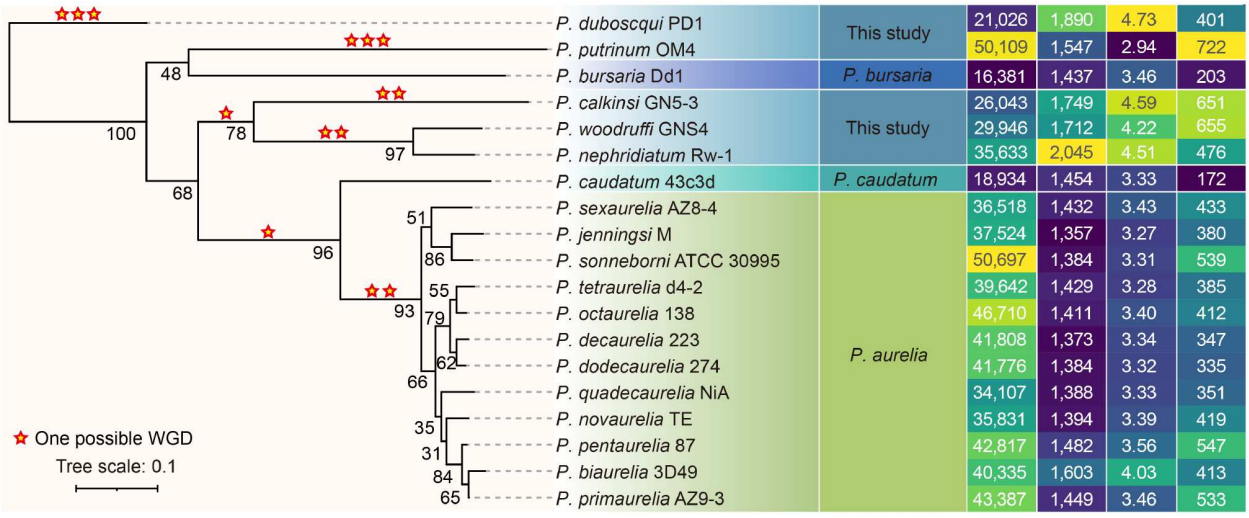

Figure 5. Phylogenomic tree of *Paramecium*. The numbers on the tree represent the bootstrap values. Stars denote possible WGD events. The heatmap on the right shows the gene count, mean coding gene length, mean exon count, and mean intergenic length for each species.

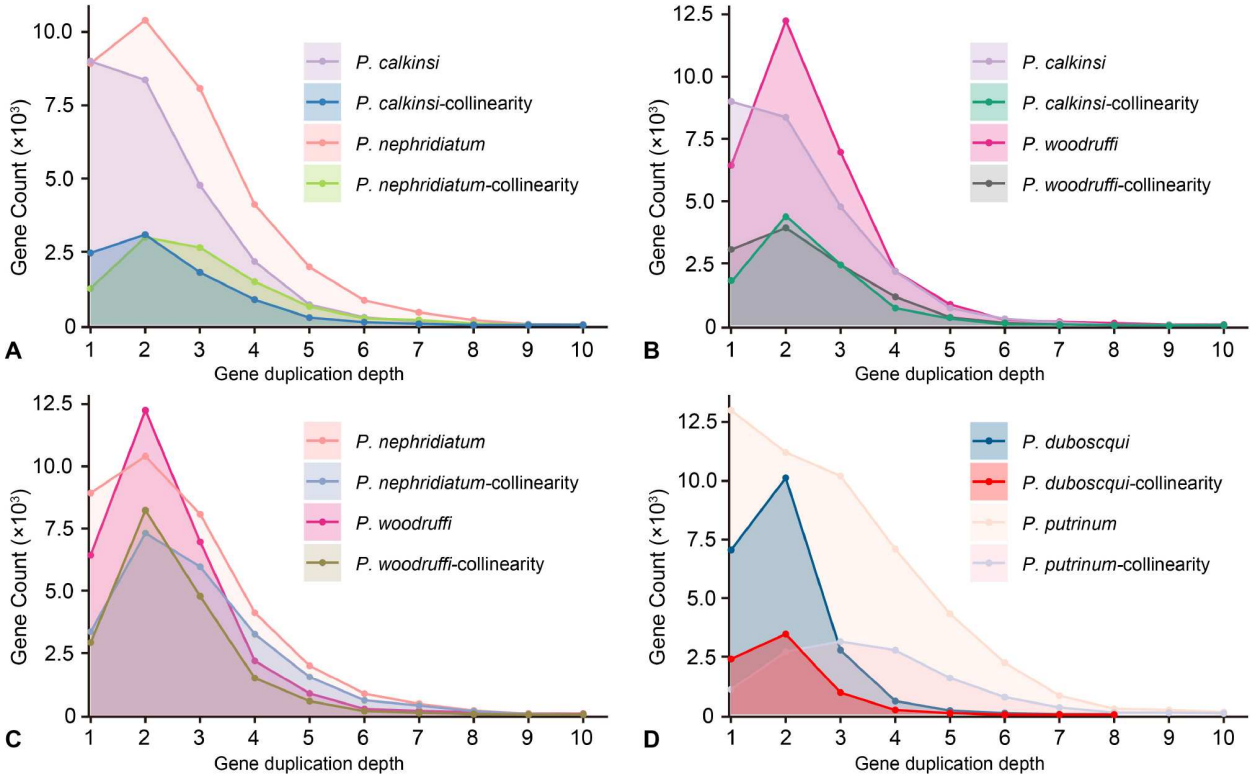

Figure 6. Duplication depth distribution. Each plot displays the gene duplication depth (the number of paralogous genes for a gene) distribution of all genes and their collinear genes. A, Relatively few collinear gene relationships between *P. calkinsi* and *P. nephridiatum*. B, Relatively few collinear gene relationships between *P. calkinsi* and *P. woodruffi*. C, High number of collinear genes between *P. woodruffi* and *P. nephridiatum*, and the collinear genes share a similar gene duplication depth distribution. D, Few collinear genes between *P. putrinum* and *P. duboscqui*, and the collinear genes have inconsistent gene duplication depth distribution.

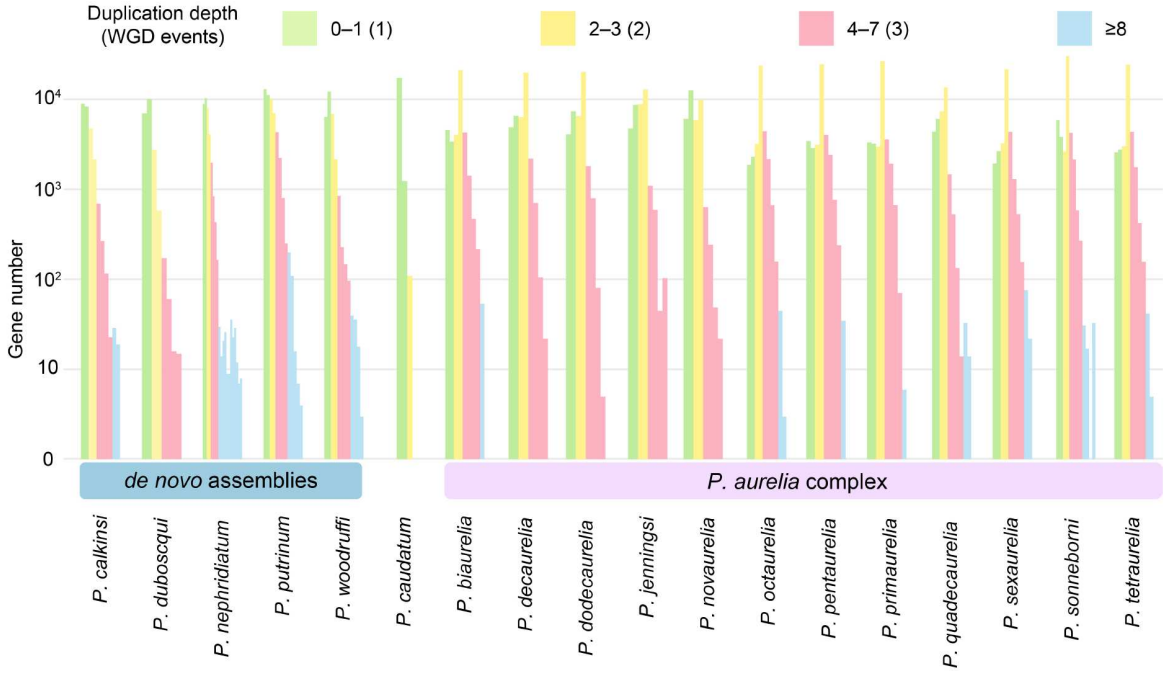

Figure 7. Distribution of gene duplication depth in different *Paramecium* species. After excluding depth with gene number <100, most species show a maximum gene duplication depth of 7 (not counting itself), matching the expected pattern of three rounds of WGDs of *P. aurelia* complex species and one round of WGD of *P. caudatum* (shown in parentheses). Five rare species thus have undergone at least three rounds of WGDs.

depths (Table S6). The results of hierarchical clustering of gene duplication depth data clearly distinguished our five species from other *Paramecium* species (Figure 8), indicating that over sufficient evolutionary time, species may exhibit similar patterns

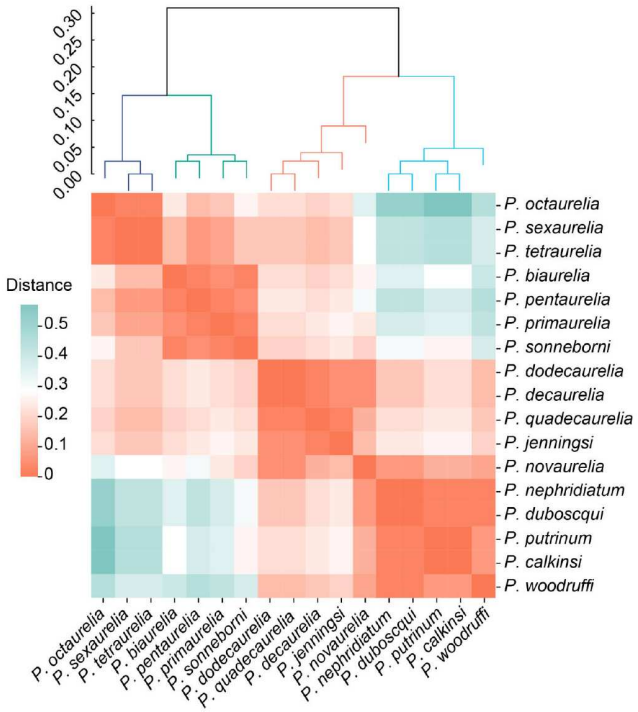

Figure 8. Spearman correlation distance and Hierarchical clustering of species based on gene duplication depth. In the heatmap, the colors closer to red indicate smaller distances and greater correlation between the species. In the hierarchical cluster tree, the coordinates represent distances.

of gene retention at various depths. The incomplete clustering of *P. aurelia* suggests that each species within this group has followed a distinct path in the evolutionary process of gradually decreasing high-duplication-depth genes, or these recently diverged species have begun to show similar patterns (Figure 8). These findings underscore the complexity and diversity of evolutionary trajectories of *Paramecium* genes after WGDs, highlighting how genomic variations and evolutionary pressures shape distinct evolutionary trajectories across *Paramecium*.

DISCUSSION

In this study, to achieve high genome completeness, we applied multiple procedures, for example, the macronuclear isolation to remove non-macronuclear cell components and bacteria in the culturing medium, PacBio HiFi sequencing, and multiple bioinformatic filters against contaminants. Despite these precautions, there was still a lack of scaffolds with two telomeres in all five *de novo* assemblies. This could be partially caused by the high sequencing error rates at numerous repetitive sequences in the genome due to the high AT content. The heterozygosity from the extreme polyploidy (840–860n) of the *Paramecium* macronuclei and no inducible autogamy could be other challenging factors for the assembling (Preer Jr, 1976; Samuel et al., 1981). The assembling of the micronuclear genomes poses even greater challenges, due to difficulties in the isolation of pure micronuclei (Guérin et al., 2017; Sellis et al., 2021). Emerging techniques, such as single-cell library preparation, flow cytometry for high-throughput nuclear sorting, as well as classical techniques such as sucrose density gradient centrifugation, may help break the technical barrier in the future (Chen et al., 2021; Guérin et al., 2017; Skoczylas and Soldo, 1975). These methods may yield high-purity micronuclei and eliminate extracellular and intracellular bacterial contaminants.

Recent research has significantly highlighted the vast genomic differentiation within the *Paramecium* genus, due to its extremely-long evolutionary history (579.5 to 1,256.4 Mya) and idiosyncratic genome architecture (Arnaiz et al., 2020; Gout et al., 2023; Kumar et al., 2017; Long et al., 2023; McGrath et al., 2014; Parfrey et al., 2011; Rataj and Vďačný, 2018). Despite similarities in morphology and life history traits, there is a substantial divergence in genome sizes. Species that have undergone the same number of WGD events even exhibit nearly two-fold differences in genome sizes. In this study, we present the macronuclear genomes of five rare species of *Paramecium*, which revealed more species-specific gene families within the genus and can provide crucial markers for species identification and primer design, facilitating further research and classification within the genus. They also confirm the universality of genome streamlining in *Paramecium*, characterized by a consistently high proportion of gene regions, with the highest of *P. caudatum* (88.94%). Such findings underscore the importance and value of broadening our research scope beyond commonly studied species to encompass more genetically diverse and rare species.

Our research demonstrates that unsynchronized WGD events and subsequent varied gene retention might have led to *Paramecium* genome divergence. Earlier studies confirmed the presence of three WGDs in *P. aurelia*, compared with its close relative, *P. caudatum*, possibly related to speciation and differentiation among *P. aurelia* sibling species and from *P. caudatum* (Gout et al., 2023; McGrath et al., 2014). Our collinearity analysis of five rare *Paramecium* indicates that non-*aurelia* congeners also experienced multiple WGDs not shared by the *P. aurelia* species. WGD events within *Paramecium* are thus more common than previously thought, not exclusive to the *P. aurelia* species complex, and have independently occurred multiple times. The median *Ks* frequency distribution of the five *de novo* assemblies shows peaks consistently at small *Ks* values, suggesting the relatively short evolutionary history of their WGDs. Nonetheless, precautions need to be taken on the application of the *Ks method* to analyze WGDs in ciliate macronuclear genomes, which usually have a very long evolutionary history, as well as extremely complicated and polyploid genome architecture. The possibility of small *Ks* values resulting from the divergence of the homologous chromosomes within the same macronuclear genome, e.g., evolution without sexual reproduction, cannot be excluded. To test such a possibility in our analyses, we analyzed whether the low *Ks* originated from homologous chromosomes heterozygosity instead of diverging paralogs. We then randomly chose two contigs (Contig5987, Contig6057) of the *P. woodruffi* assembly, with low *Ks* values and abundant collinear blocks between each other. Despite the low number of synonymous nucleotide substitutions, we detected significant differences between collinear blocks, especially in the gene content and length, intron count, etc., which cannot be structural variations from mutation accumulation in homologous chromosomes over the generally short time span after the last sexual process (Figure 9; Tables S7–S10). Together with the collinearity analyses, the *Ks* method thus reliably revealed the WGDs in the five rare species.

Our phylogenetic tree, constructed from whole-genome sequences of *Paramecium* exhibits high confidence levels at the nodes across different clades. However, some nodes show lower

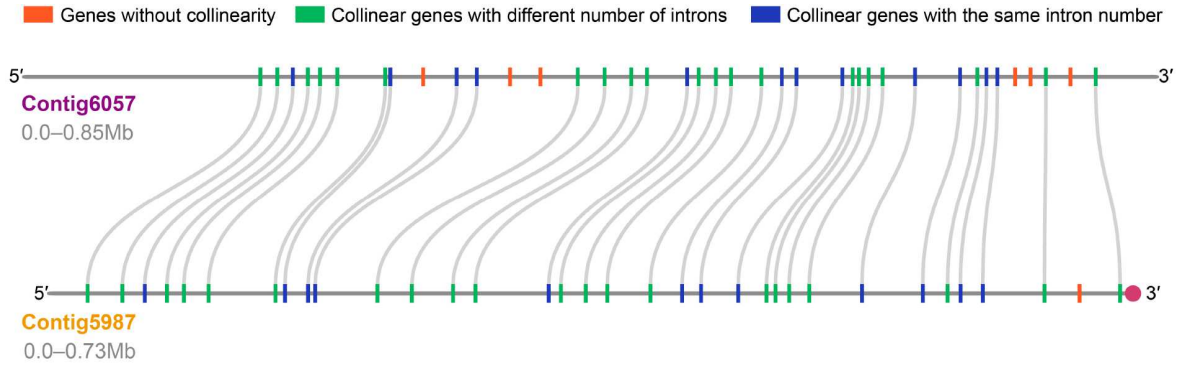

Figure 9. Details of collinearity between two contigs of *P. woodruffi* for demonstration. Grey lines connect genes with collinearity. Contig5987 has telomere sequences in its 3′ end (pink circle).

confidence, likely due to the substantial diversity within *Paramecium*. Based on current genomic data, *P. bursaria*, *P. duboscqui*, and *P. putrinum* are all potential basal species within the genus, as previously suggested by ribosomal subunit sequence phylogeny (Boscaro et al., 2012; Fokin and Chivilev, 2000; Strüder-Kypke et al., 2000). Regarding the *P. woodruffi* group, *P. woodruffi* and *P. nephridiatum* are closely related, whereas *P. calkinsi* shows greater genomic divergence from these two. Morphologically, *P. woodruffi* was considered to have given rise to other species within the group (Fokin and Chivilev, 2000). Now phylogenetic relationship, gene family distribution, and collinearity relationship based on whole genome data suggest that *P. calkinsi* may be closer to the ancestor of the narrow *P. woodruffi* group (contains *P. woodruffi* and *P. nephridiatum*). Another contentious issue in *Paramecium* is the evolutionary status of *P. polycaryum*. Unfortunately, a complete genome for *P. polycaryum* is currently unavailable, which hinders the determination of its phylogenetic position at a whole-genome scale. Aside from the taxonomic status between species, the question of whether each subgenus/species within the *Paramecium* genus contains cryptic species has not been fully answered (Schrallhammer et al., 2006; Strüder-Kypke et al., 2000; Tarcz et al., 2014). Morphological features, mating experiments, and genomic studies have strongly supported the existence of the *P. aurelia* species complex (Gout et al., 2023; Sonneborn, 1975; Wichterman, 2012). While in most other species, only DNA fragments and symbiont types have been used to identify potential subtype relationships among different strains of the same species (Schrallhammer et al., 2006; Strüder-Kypke et al., 2000; Tarcz et al., 2014). There is still a lack of detailed mating experiments or whole genome data to support these findings.

Together with previously published *Paramecium* macronuclear genomes, the *de novo* assemblies of the five rare species give us a valuable opportunity to assemble a pan-genome that could aid future research. By merging or splitting chromosomes, we used clustering algorithms such as greedy algorithms and maximum clique algorithms to select chromosomes with high similarity for alignment. Despite the low proportion of unique genes in each *Paramecium* species, aligning their chromosomes is difficult. This makes generating clear cactus graphs challenging, likely due to significant variations in chromosome number and length, as well as differences in gene order and composition between species. This also limits our further acquisition of the accessory genomes and core genomes of the *Paramecium* genus. Despite these obstacles, the effort to assemble a *Paramecium* pan-genome is not without potential. The integration of more advanced computational approaches could substantially improve our ability to tackle the complexities inherent in the *Paramecium* pan-genome. Furthermore, expanding the genomic database with more extensive sampling from various ecological niches and geographic locations would enhance the representativeness of the pan-genome, allowing for a more accurate reflection of *Paramecium* genetic diversity and evolution. This broader genomic representation would also facilitate the discovery of novel genes and alleles, providing deeper insights into the ecological roles and evolutionary pressures shaping this genus.

This study significantly enhances the understanding of the *Paramecium* genus by delving into the genomes of five rare species, revealing substantial genomic diversity and divergence not previously captured. These findings highlight the widespread occurrences of multiple whole genome duplications (WGDs), which were thought to be only present in the *P. aurelia* species complex. The phylogenomic analyses by the inclusion of the five *de novo* assemblies have also successfully resolved evolutionary relationships that were previously unattainable due to insufficient data, such as the existence of the *P. woodruffi*-group and *P. duboscqui* being the basal species of *Paramecium*. Genomic resources from rare species can thus reveal unknown genome diversity and assist in resolving uncertainties in microbial eukaryote studies.

MATERIALS AND METHODS

Strains and media

Strain details are in Table S11. *Paramecium* cultures were maintained in a lettuce-based medium prepared using the following method: lettuce leaves were washed in deionized water and then briefly boiled for 1–2 min. After boiling, the leaves were immediately cooled in ice-cold deionized water. The cooled leaves were then juiced, and the juice was filtered through eight layers of gauze. For every kilogram of lettuce leaf, the juice was diluted to a total volume of 2 liters with deionized water. A calcium-free modified Dryl's solution was then prepared by dissolving 88.2 g of trisodium citrate dihydrate ($C_6H_5Na_3O_7 \cdot 2H_2O$), 75.2 g of disodium hydrogen phosphate dihydrate ($NaH_2PO_4 \cdot 2H_2O$), and 12.3 g of potassium dihydrogen phosphate (KH_2PO_4) in 3 L of deionized water.

The final medium was composed of 40 mL of the prepared lettuce juice, 32 mL of the calcium-free Dryl's solution, and 1,516 mL of deionized water, to which 12 mL of a 200 mmol L^{-1} calcium chloride ($CaCl_2$) solution was added. The medium was inoculated with *Klebsiella pneumoniae* 6081 and incubated at 24°C–25°C for 24 h.

Nucleic acid extraction, macronuclear isolation, and sequencing

After 5–6 days' incubation in 1 L lettuce-based medium, *Paramecium* cells were concentrated through centrifugation and re-suspended in 100 μL Dryl's solution. Half of the re-suspended cells were subjected to RNA extraction using TRIzol (Invitrogen, USA). The aqueous supernatant from TRIzol/chloroform treatment was precipitated with isopropanol, and the pellet was washed in 70% ethanol before suspension in H_2O. RNA was cleaned up using the RNeasy Mini Kit (Qiagen, Germany).

The rest of the concentrated cells were used for the isolation of macronuclei. The cells for macronuclear isolation were initially rinsed to remove residual medium using a solution composed of 10 mmol L^{-1} Tris (pH 7.2), 0.25 mol L^{-1} sucrose, and 2 mmol L^{-1} $MgCl_2$, through centrifugation at 100 g for 2 min. Following the rinsing, cells were lysed using a lysis solution containing 10 mmol L^{-1} Tris (pH 7.2), 0.25 mol L^{-1} sucrose, 2 mmol L^{-1} $MgCl_2$, 0.5% Nonidet P40, and 0.1% sodium deoxycholate. The lysis mixture was then centrifuged at 100 g for 2 min. After removing the supernatant, the lysis solution was added once more, followed by incubation on ice for 10–30 min and another centrifugation at 100 g for 2 min, resulting in white macronuclear precipitates. Macronuclear genomic DNA was then extracted using the MasterPure Complete DNA and RNA Purification Kit (Lucigen, USA).

For sequencing, macronuclear genomic DNA >60 kb and DIN scores above 9 were prepared by MagAttract HMW DNA Kit (Qiagen). DNA samples were then sent to UC Berkeley QB3 Genomics for PacBio HiFi sequencing. RNA was extracted using a TRIzol-based method to ensure minimal degradation. Samples achieving an RNA Integrity Number (RIN) of 4–8 were sent to TGen for RNA-seq sequencing (Illumina PE150). Since *P. nephridiatum* Rw-1 is a slow-growing strain with repeated failed RNAseq trials, we then prepared RNAseq libraries using NEBNext® Single Cell/Low Input RNA Library Prep Kit for Illumina® (New England Biolabs, USA) and then performed Illumina PE150 sequencing at TGen.

Genome assembly

Draft genomes were assembled by Canu v2.2 from HiFi reads (Koren et al., 2017). Redundant scaffolds were subsequently filtered twice by purge_dups v1.2.5 (Guan et al., 2020). To remove bacterial contaminant sequences, the purged genomes were screened using BLASTN v2.13.0+ against a bacterial genome database, and scaffolds with more than 50% bacterial content were discarded.

To assess scaffold quality, Quast v5.2.0 was run to generate GC content distribution plots for each scaffold (Gurevich et al., 2013). Based on the GC peaks observed in these plots, scaffolds with exceptionally high GC content were removed. *Paramecium* 18S rDNA-containing scaffolds typically show a significantly higher GC content than the genome average. To prevent erroneous removal, scaffolds with high GC content were analyzed using RNAmmer v1.2 to identify 18S rDNA (Lagesen et al., 2007). Further confirmation of these scaffolds with 18S rRNA sequences was done by blasting them against known *Paramecium* 18S rDNA sequences. Additionally, scaffolds that were filtered out but contained telomeric sequences were recovered and aligned against the retained scaffolds using CD-HIT v4.8.1 (Li and Godzik, 2006). All duplicate sequences were removed, and unique scaffolds with telomeric sequences were preserved. This approach avoided the loss of crucial genomic elements due to over-filtering, thereby maintaining the integrity of the genome assembly.

Structural and functional annotation

RNAseq reads were first processed using fastp v0.20.0 (Chen et al., 2018), with Phred quality scores >15 or higher being retained. Following quality control, these reads were mapped to the assembled genomes using HISAT2 v2.2.1 (Kim et al., 2019). The aligned reads were then used to assemble the transcriptome by Trinity v2.12.0 (Grabherr et al., 2011). Structural annotation was conducted using EuGene v4.3, with a Weight Array Matrix (WAM) built from annotation files of all species published in ParameciumDB (Arnaiz et al. 2020; Sallet et al. 2019). Due to *Paramecium*'s unique codon usage, protein sequences were transformed using a modified script in AUGUSTUS, named gtf2aa.pl (Stanke et al., 2006). Genes that did not end with the *Paramecium*-specific stop codon TAG or those containing multiple TGA codons within the coding sequence were identified and excluded from both the protein sequence and GFF3 files. The quality of the annotations was evaluated using BUSCO v5.4.7 with the parameters: -m proteins -l alveolata_odb10 (Manni et al., 2021). Functional annotation was performed using OmicsBox (https://www.biobam.com/omicsbox/). The structural information of other *Paramecium* genomes (Tables S2 and S4) was analyzed using a modified lipm_genome_statistics.pl script within EuGene for annotation descriptions. After mapping the RNASeq data to the transcriptome using HISAT2 v2.2.1, we identified alternative splicing events using rMATS v4.1.1, with default parameters. Among the five species analyzed, alternative splicing events were only detected in *P. nephridiatum* (32 events, less than 0.09% of the total genes involved).

Comparative genomics

Gene alignments from BLASTP were for exploring collinear relationships and gene duplication depth using MCScanX (Wang et al., 2012), limiting to collinear blocks containing more than five genes for clarity. To analyze the *Ks*, we ran WGDI v0.6.4, which features an optimized collinearity algorithm (Sun et al., 2022). Due to the inherently short chromosomes in *Paramecium*, many chromosomes were excluded during the *Ks* peak calculation by the program. We randomly chose two collinear scaffolds, parsed their collinearity gene, intron, and protein sequences, and conducted one-to-one identity and length comparison using BLAST v2.13.0+ and SeqKit v2.2.0 (Altschul et al., 1990; Shen et al., 2016), revealing the large-scale indels hidden behind the low *Ks*.

Data visualization

The genome Circos plot was done by the Advanced Circos plugin

in TBtools v2.080 (Chen et al., 2023). The order of the chromosomes in the circos plot was rearranged using an annealing algorithm, to make most of the collinear chromosomes distribute in the same one-third of the circle. We used OrthoFinder v2.5.5 to identify orthologous genes and build the phylogenetic tree (Emms and Kelly, 2019). We used custom Python scripts with Biopython, ete3, and Bio. Phylo library to identify monophyletic groups and calculate the Robinson-Foulds distance between the gene tree and the reference tree. A Venn plot is generated in EVenn (Yang et al., 2024). All other figures and statistics were done in R v4.3.2 (R Core Team, 2023) or Python scripts.

Hierarchical cluster

Hierarchical clustering is an unsupervised machine learning technique for grouping data points based on similarities and distances, and it can better reflect the natural structure of the data (Hastie et al., 2009; James et al., 2013). As small-scale duplications or horizontal gene transfers usually involve only a few genes, we excluded the duplication depths with gene counts fewer than 100 from the five *de novo* assemblies. This resulted in a maximum duplication depth of around 7, suggesting that three WGD events occurred. We then filtered out genes with depths greater than 7 (8 genes if counting itself), corresponding to the theoretical depth limits of three consecutive WGD events (2^3). Subsequently, we normalized the remaining data to obtain the relative proportions of genes at each depth as input data for calculating the cluster distances. Since the distribution of gene duplication depth data is unknown, we chose the non-parametric method—Spearman correlation distance (eq. 1)—to compute the distance or similarity between each pair of species based on the normalized gene duplication depth data.

$$d(X, Y) = \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)}. \quad (1)$$

Here, $d_i = rank(X_i) - rank(Y_i)$, which represents the difference in paired ranks, and n is the number of cases, which is 8 of our data. And d_i refers to the ranked difference between the ith gene duplication depths of paired species 8. The Spearman correlation distance matrix is shown in Figure 8. Then, we proceed with the Hierarchical cluster in the following steps:

(i) Start by assigning each species to a cluster, so we have 18 clusters, each containing one species. Let the distances between the clusters be the same as the distances between each species.

(ii) Use the average linkage method (eq. 2) to find the closest pair of clusters and merge them into a single cluster, so now we have 17 clusters.

The formula for the distance $d(A, B)$ between clusters A and B is given by

$$d(A, B) = \frac{1}{n_A n_B} \sum_{k \in A} \sum_{l \in B} d_{kl}, \quad (2)$$

and d_{kl} are the Spearman correlation distances between each species calculated by eq. (1).

(iii) Compute distances between the new cluster and each of the old clusters.

(iv) Repeat steps 2 and 3 until all species are clustered into a single cluster of size 18 (Figure 8).

Data and materials availability

Genomes of this study were uploaded to the China National Center for Bioinformation with project number PRJCA025782. The HiFi reads of the five rare *Paramecium* were uploaded to SRA with the Bioproject number PRJNA1107025. Strain cultures of the five species can be accessed through http://nbrpcms.nig.ac.jp/paramecium/strain/?lang=e with NBRP ID of PK000001A (*P. calkinsi* GN5-3), PD000001A (*P. duboscqui* PD1), PN000001A (*P. nephridiatum* Rw-1), PP004001A (*P. putrinum* OM4) and PW000001A (*P. woodruffi* GNS4).

Compliance and ethics

The authors declare that they have no conflict of interest.

Acknowledgement

This work was supported by the Laoshan Laboratory (LSKJ202203203), the National Natural Science Foundation of China (31961123002, 32270435 and 32471688), the National Institutes of Health (R35-GM122566-01) and the National Science Foundation (DBI-2119963, DEB-1927159 and 1911449). We appreciate the technical help of Xianyu Yang, Masahiro Fujishima, and Angelica Urquidez.

Supporting information

The supporting information is available online at https://10.1007/s11427-024-2872-7. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol 215, 403–410.

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

Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444, 171–178.

Beisson, J., and Sonneborn, T.M. (1965). Cytoplasmic inheritance of the organization of the cell cortex in *Paramecium aurelia*. Proc Natl Acad Sci USA 53, 275–282.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., Da Silva, C., Labadie, K., Alberti, A., et al. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun 5, 1–10.

Bondarenko, V.S., Gelfand, M.S., and Rogozin, I.B. (2016). Evolution of the exon-intron structure in ciliate genomes. PLoS One 11, e0161476.

Boscaro, V., Fokin, S.I., Verni, F., and Petroni, G. (2012). Survey of *Paramecium duboscqui* using three markers and assessment of the molecular variability in the genus *Paramecium*. Mol Phylogenet Evol 65, 1004–1013.

Catania, F., Rothering, R., Vitali, V., and Zufall, R. (2021). One cell, two gears: extensive somatic genome plasticity accompanies high germline genome stability in *Paramecium*. Genome Biol Evol 13, evab263.

Catania, F., Wurmser, F., Potekhin, A.A., Przybos, E., and Lynch, M. (2008). Genetic diversity in the *Paramecium aurelia* species complex. Mol Biol Evol 26, 421–431.

Chang, N., Sun, Q., Hu, J., An, C., and Gao, H. (2017). Large introns of 5 to 10 kilo base pairs can be spliced out in *Arabidopsis*. Genes 8, 200.

Chen, C., Wu, Y., Li, J., Wang, X., Zeng, Z., Xu, J., Liu, Y., Feng, J., Chen, H., He, Y., et al. (2023). TBtools-II: a "one for all, all for one" bioinformatics platform for biological big-data mining. Mol Plant 16, 1733–1742.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890.

Chen, W., Zuo, C., Wang, C., Zhang, T., Lyu, L., Qiao, Y., Zhao, F., and Miao, M. (2021). The hidden genomic diversity of ciliated protists revealed by single-cell genome sequencing. BMC Biol 19, 264.

Cheng, Y.H., Liu, C.F.J., Yu, Y.H., Jhou, Y.T., Fujishima, M., Tsai, I.J., and Leu, J.Y. (2020). Genome plasticity in *Paramecium bursaria* revealed by population genomics. BMC Biol 18, 180.

Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20, 1–14.

Fokin, S.I. (2010). *Paramecium* genus: biodiversity, some morphological features and the key to the main morphospecies discrimination. Protistology, 6: 227–235.

Fokin, S.I., and Chivilev, S.M. (2000). *Paramecium* morphometric analysis and taxonomy. Acta Protozool, 39: 1–14.

Fokin, S.I., Przybos, E., and Chivilev, S.M. (2001). Nuclear reorganization variety in *Paramecium* (Ciliophora: Peniculida) and its possible evolution. Acta Protozool, 40: 249–262.
Fokin, S.I., Stoeck, T., and Schmidt, H.J. (1999). Rediscovery of *Paramecium nephridiatum* gelei, 1925 and its characteristics. J Eukaryot Microbiol 46, 416–426.
Garnier, O., Serrano, V., Duharcourt, S., and Meyer, E. (2004). RNA-mediated programming of developmental genome rearrangements in *Paramecium tetraurelia*. Mol Cell Biol 24, 7370–7379.
Gout, J.F., Hao, Y., Johri, P., Arnaiz, O., Doak, T.G., Bhullar, S., Couloux, A., Guérin, F., Malinsky, S., Potekhin, A., et al. (2023). Dynamics of gene loss following ancient whole-genome duplication in the cryptic *Paramecium* complex. Mol Biol Evol 40, msad107.
Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29, 644–652.
Greczek-Stachura, M., Rautian, M., and Tarcz, S. (2021). *Paramecium bursaria*—a complex of five cryptic species: mitochondrial DNA COI haplotype variation and biogeographic distribution. Diversity 13, 589.
Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y., Durbin, R., and Valencia, A. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics 36, 2896–2898.
Guérin, F., Arnaiz, O., Boggetto, N., Denby Wilkes, C., Meyer, E., Sperling, L., and Duharcourt, S. (2017). Flow cytometry sorting of nuclei enables the first global characterization of *Paramecium* germline DNA and transposable elements. BMC Genomics 18, 327.
Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075.
Hastie, T., Tibshirani, R., and Friedman, J.H. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Vol 2 (New York: Springer).
James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning, Vol 112 (New York: Springer).
Jankowski, A. (1972). Cytogenetics of *Paramecium purtinum* C. et L., 1858. Acta Protozool 10, 285–394.
Jin, D., Li, C., Chen, X., Byerly, A., Stover, N.A., Zhang, T., Shao, C., and Wang, Y. (2023). Comparative genome analysis of three euplotid protists provides insights into the evolution of nanochromosomes in unicellular eukaryotic organisms. Mar Life Sci Technol 5, 300–315.
Johri, P., Gout, J.F., Doak, T.G., Lynch, M., and Wittkopp, P. (2022). A population-genetic lens into the process of gene loss following whole-genome duplication. Mol Biol Evol 39, msac118.
Johri, P., Krenek, S., Marinov, G.K., Doak, T.G., Berendonk, T.U., and Lynch, M. (2017). Population genomics of *Paramecium* species. Mol Biol Evol 34, 1194–1216.
Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37, 907–915.
Kimball, R.F. (1943). Mating types in the ciliate protozoa. Q Rev Biol 18, 30–45.
Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res 27, 722–736.
Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol 34, 1812–1819.
Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.H., Rognes, T., and Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35, 3100–3108.
Li, C., Chen, X., Zheng, W., Doak, T.G., Fan, G., Song, W., and Yan, Y. (2021). Chromosome organization and gene expansion in the highly fragmented genome of the ciliate *Strombidium stylifer*. J Genet Genomics 48, 908–916.
Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.
Li, Z., Tiley, G.P., Galuska, S.R., Reardon, C.R., Kidder, T.I., Rundell, R.J., and Barker, M.S. (2018). Multiple large-scale gene and genome duplications during the evolution of hexapods. Proc Natl Acad Sci USA 115, 4713–4718.
Long, H., Doak, T.G., and Lynch, M. (2018). Limited mutation-rate variation within the *Paramecium aurelia* species complex. G3 Genes Genomes Genet 8, 2523–2526.
Long, H., Johri, P., Gout, J.F., Ni, J., Hao, Y., Licknack, T., Wang, Y., Pan, J., Jiménez-Marín, B., and Lynch, M. (2023). *Paramecium* genetics, genomics, and evolution. Annu Rev Genet 57, 391–410.
Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science 290, 1151–1155.
Lyu, L., Zhang, X., Gao, Y., Zhang, T., Fu, J., Stover, N.A., and Gao, F. (2024). From germline genome to highly fragmented somatic genome: genome-wide DNA rearrangement during the sexual process in ciliated protists. Mar Life Sci Technol 6, 31–49.
Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., Zdobnov, E.M., and Kelley, J. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol 38, 4647–4654.
McGrath, C.L., Gout, J.F., Doak, T.G., Yanagi, A., and Lynch, M. (2014). Insights into three whole-genome duplications gleaned from the *Paramecium caudatum* genome sequence. Genetics 197, 1417–1428.
Mouillot, D., Bellwood, D.R., Baraloto, C., Chave, J., Galzin, R., Harmelin-Vivien, M., Kulbicki, M., Lavergne, S., Lavorel, S., Mouquet, N., et al. (2013). Rare species support vulnerable functions in high-diversity ecosystems. PLoS Biol 11, e1001569.
Parfrey, L.W., Lahr, D.J.G., Knoll, A.H., and Katz, L.A. (2011). Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc Natl Acad Sci USA 108, 13624–13629.
Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, M.C., and Caracausi, M. (2019). Human protein-coding genes and gene feature statistics in 2019. BMC Res Notes 12, 315.
Potekhin, A., and Mayén-Estrada, R. (2020). *Paramecium* diversity and a new member of the *Paramecium aurelia* species complex described from Mexico. Diversity 12, 197.
Preer Jr, J.R. (1976). Quantitative predictions of random segregation models of the ciliate macronucleus. Genet Res 27, 227–238.
Przyboś, E., Rautian, M., Beliavskaia, A., and Tarcz, S. (2019). Evaluation of the molecular variability and characteristics of *Paramecium polycaryum* and *Paramecium nephridiatum*, within subgenus *Cypriostomum* (Ciliophora, Protista). Mol Phylogenet Evol 132, 296–306.
R Core Team, R. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from URL: https://www.R-project.org/.
Rataj, M., and Vďačný, P. (2018). Dawn of astome ciliates in light of morphology and time-calibrated phylogeny of *Haptophrya planariarum*, an obligate endosymbiont of freshwater turbellarians. Eur J Protistol 64, 54–71.
Sabaneyeva, E. (1997). Extrachromosomal nucleolar apparatus in the macronucleus of the ciliate *Paramecium putrinum*: LM, EM and confocal microscopy studies. Arch Protistenk 148, 365–373.
Sallet, E., Gouzy, J., and Schiex, T. (2019). EuGene: an automated integrative gene finder for eukaryotes and prokaryotes. In Pollen: Gene Prediction: Methods and Protocols, M. Kollmar, ed. (New York, Springer New York), pp. 97–120.
Samuel, C., Mackie, J., and Sommerville, J. (1981). Macronuclear chromatin organization in *Paramecium primaurelia*. Chromosoma 83, 481–492.
Schrallhammer, M., Fokin, S.I., Schleifer, K., and Petroni, G. (2006). Molecular characterization of the obligate endosymbiont "*Caedibacter macronucleorum*" Fokin and Görtz, 1993 and of its host *Paramecium duboscqui* strain Ku4-8. J Eukaryot Microbiol 53, 499–506.
Sellis, D., Guérin, F., Arnaiz, O., Pett, W., Lerat, E., Boggetto, N., Krenek, S., Berendonk, T., Couloux, A., Aury, J.M., et al. (2021). Massive colonization of protein-coding exons by selfish genetic elements in *Paramecium* germline genomes. PLoS Biol 19, e3001309.
Session, A.M., Uno, Y., Kwon, T., Chapman, J.A., Toyoda, A., Takahashi, S., Fukui, A., Hikosaka, A., Suzuki, A., Kondo, M., et al. (2016). Genome evolution in the allotetraploid frog *Xenopus laevis*. Nature 538, 336–343.
Shen, W., Le, S., Li, Y., Hu, F., and Zou, Q. (2016). SeqKit: a cross-platform and Ultrafast Toolkit for FASTA/Q file manipulation. PLoS One 11, e0163962.
Skoczylas, B., and Soldo, A.T. (1975). Separation and purification of macronuclei from macronuclear fragments and micronuclei in the ciliate *Paramecium aurelia*. Exp Cell Res 90, 143–152.
Sonneborn, T.M. (1937). Sex, sex inheritance and sex determination in *Paramecium Aurelia*. Proc Natl Acad Sci USA 23, 378–385.
Sonneborn, T.M. (1970). Chapter 12 Methods in Paramecium Research. In Pollen: Methods in Cell Biology, D.M. Prescott, ed. (New York: Academic Press), pp. 241–339.
Sonneborn, T.M. (1975). The*Paramecium aurelia* complex of fourteen sibling species. Trans Am Microscopical Soc 94, 155–178.
Spingola, M., Grate, L., Haussler, D., and Ares, M. (1999). Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. RNA 5, 221–234.
Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Res 34, W435–W439.
Strüder-Kypke, M.C., Wright, A.D.G., Fokin, S.I., and Lynn, D.H. (2000). Phylogenetic relationships of the genus *Paramecium* inferred from small subunit rRNA gene sequences. Mol Phylogenet Evol 14, 122–130.
Sun, P., Jiao, B., Yang, Y., Shan, L., Li, T., Li, X., Xi, Z., Wang, X., and Liu, J. (2022). WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. Mol Plant 15, 1841–1851.
Tarcz, S., Rautian, M., Potekhin, A., Sawka, N., Beliavskaya, A., Kiselev, A.,

Nekrasova, I., and Przyboś, E. (2014). *Paramecium putrinum* (Ciliophora, Protozoa): the first insight into the variation of two DNA fragments—molecular support for the existence of cryptic species. Mol Phylogenet Evol 73, 140–145.
Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40, e49.
Wichterman, R. (2012). The Biology of *Paramecium* (New York: Springer US).
Woodruff, L.L. (1921). The structure, life history, and intrageneric relationships of *Paramecium calkinsi*, sp. nov. Biol Bull 41, 171–180.
Yan, Y., Maurer-Alcalá, X.X., Knight, R., Kosakovsky Pond, S.L., Katz, L.A., and Cavanaugh, C.M. (2019). Single-cell transcriptomics reveal a correlation between genome architecture and gene family evolution in ciliates. mBio 10, e02524-19.
Yang, M., Chen, T., Liu, Y.H., and Huang, L. (2024). Visualizing set relationships: EVenn's comprehensive approach to Venn diagrams. iMeta 3, e184.
Zhang, X., Zhao, Y., Zheng, W., Nan, B., Fu, J., Qiao, Y., Zufall, R.A., Gao, F., and Yan, Y. (2023). Genome-wide identification of ATP-binding cassette transporter B subfamily, focusing on its structure, evolution and rearrangement in ciliates. Open Biol 13, 230111.
Zheng, W., Chen, J., Doak, T.G., Song, W., Yan, Y., and Birol, I. (2020). ADFinder: accurate detection of programmed DNA elimination using NGS high-throughput sequencing data. Bioinformatics 36, 3632–3636.
Zheng, W., Wang, C., Lynch, M., Gao, S., Katz, L.A., and Capone, D.G. (2021). The compact macronuclear genome of the ciliate *Halteria grandinella*: a transcriptome-like genome with 23,000 nanochromosomes. mBio 12, e01964-20.