GBE

De Novo Genome Sequence Assembly of the RNAi-Tractable *Paramecium bursaria* 186b: An Endosymbiotic Model System

Guy Leonard [1], Benjamin H. Jenkins [1,2], Fiona R. Savory [1], Estelle S. Kilias [1], Finlay Maguire [3,4], David S. Milner [1], Thomas A. Richards [1,*]

[1]Department of Biology, University of Oxford, Oxford OX1 3SZ, UK

[2]Department of Biochemistry, University of Cambridge, Cambridge CB2 1QW, UK

[3]Department of Community Health and Epidemiology, Dalhousie University, Centre for Clinical Research, Halifax, NS B3H 1V7, Canada

[4]Faculty of Computer Science, Dalhousie University, Halifax, NS B3H 3P8, Canada

*Corresponding author: E-mail: thomas.richards@biology.ox.ac.uk.

Accepted: August 12, 2025

Abstract

How two species engage in stable endosymbiosis is a biological quandary. The study of facultative endosymbiotic interactions has emerged as a useful approach to understand how endosymbiotic functions can arise. The ciliate protist *Paramecium bursaria* hosts green algae of the order Chlorellales in a facultative photo-endosymbiosis. We have recently reported RNAi as a tool for understanding gene function in *P. bursaria* 186b (CCAP strain 1660/18). To complement this work, here we report a near complete host genome and transcriptome sequence dataset, using both Illumina and PacBio sequencing methods, in order to aid genome analysis and to enable the design of RNAi experiments. Our analyses demonstrate *P. bursaria* 186b, like other ciliates such as diverse species of *Paramecia*, possess numerous tiny introns. These data patterns, combined with the alternative genetic code common to ciliates, make gene identification and annotation challenging; as such, we identify gene models using Iso-Seq methodologies. These data will aid the investigation of genome evolution in the *Paramecia* and provide additional source data for the exploration of endosymbiotic functions.

Key words: ciliates, endosymbiosis, introns, codon usage.

Significance

How two species engage in endosymbiosis, with one cell living within the other, is a biological quandary. *Paramecium bursaria* is a single-celled protist which hosts hundreds of green algae in a nascent photo-endosymbiosis. Here we report host genome and transcriptome sequence datasets of one strain of *P. bursaria.* Combined with development of reverse genetic methodologies for this strain, these data will enable the design of genetic experiments to manipulate host function. The genome assembly will therefore aid the investigation of genome evolution in the *Paramecia* and provide additional source data for the exploration of endosymbiotic functions.

Introduction

Endosymbiosis is a key phenomenon which has played an important role in the early evolution of eukaryotic cellular complexity (Bonen and Doolittle 1975; Bonen et al. 1977) and the diversification of eukaryotic forms from algae to corals to insects (e.g. Archibald (2009); Curtis et al. (2012); Keeling (2013); Kwong et al. (2019); McCutcheon et al. (2024)). *Paramecium bursaria* (*Pb*) is a ciliate protist

and a member of the Alveolata supergroup (Adl et al. 2019). Like all ciliates, *Pb* possesses two nuclei: a macronucleus which encodes somatic function and is typically characterized by short chromosomes with a high ploidy count and a transcriptionally inactive diploid micronucleus which engages in infrequent sexual reproduction (Aury et al. 2006). This single-celled organism, in its natural state, hosts in excess of 100 green algae in a stable but facultative endosymbiosis (Siegel 1960). Cell sampling, culturing, and rDNA marker sequencing combined with phylogenetics have shown that the *Pb* species complex is composed of numerous "syngens" (i.e. complementary mating type groups) with variant biogeographical provenance and which may represent cryptic species (Greczek-Stachura et al. 2012; Spanner et al. 2022).

The *Pb* system has emerged as a powerful model system for conducting experimental research on how two distinct organisms' function within an endosymbiotic interaction (e.g. Siegel (1960); Karakashian and Karakashian (1965); Kato and Imamura (2009); Kodama and Fujishima (2013); Lowe et al. (2016)). As a photosymbiotic protist, *Pb* cultures are easy to grow, and several characteristics of the endosymbiotic interaction are directly observable using microscopy. These characteristics include, for example, the chlorophyll status of *Pb* cells (a proxy for the wider status of the algal population) (Jenkins et al. 2021a). Important work has also shown that the endosymbiosis is based on algal secretion of photosynthesis-derived fixed carbon in the form of the dissacharide maltose. This behavior is triggered by moderately acidic pH conditions, a likely outcome induced when the algae become enclosed within the host phagotrophic-derived symbiosome, known as the perialgal vacuole (Muscatine et al. 1967; Kessler et al. 1991; Kato and Imamura 2009). Much of the experimental progress for this system is underpinned by the capacity to separate host and endosymbiont, culture the partners separately, and then re-initiate the endosymbiosis (Kessler et al. 1991; Kato and Imamura 2009). The capacity to separate the partners has been used to explore compatibility between different strains of host and endosymbiont, demonstrating variant interaction responses (Bomford 1965; Reisser 1987; Takeda et al. 1998; Minter et al. 2018; Sørensen et al. 2021). This indicates that variant syngens have distinct genetic, phenotypic, and endosymbiotic interaction characteristics.

Our long-term aim is to develop *Pb* 186b as a model organism for studying the cell biology of facultative phototrophic endosymbiotic interactions in order to understand how cellular mechanisms that support endosymbiotic interactions evolved to allow for interaction stability. To this end, we, and others, have developed RNA interference (RNAi) gene knockdown methods which can enable rapid assessment of gene functions which control endosymbiotic interactions (He et al. 2019; Jenkins et al. 2021a, 2021b). We found *Pb* 186b (CCAP 1660/18) (Spanner et al. 2020, 2022) is readily tractable for RNAi and suspect that specific culture conditions would make other strains also RNAi tractable. We note that sequencing initiatives have produced draft genome assemblies for five additional strains of *Pb* (Kodama and Fujishima 2013; He et al. 2019; Cheng et al. 2020). To further facilitate the use of *Pb* 186b for functional experimentation, we report the draft macronuclear genome assembly and annotation of this RNAi-inducible strain. Genome annotation of ciliates such as *Paramecium* with a macronuclear chromosome structure, non-universal genetic code (Prescott 1994) and tiny introns (Russell et al. 1994; Jaillon et al. 2008) is a significant challenge. In response to this challenge we have sequenced the transcriptome using Iso-Seq methods.

Results and Discussion

Genome Assembly and Annotation

Here, we have generated genome and transcriptome sequencing data using PacBio and Illumina technologies for *Pb* 186b (CCAP 1660/18). The *Pb* culture represents a consortium of the host ciliate, one or more species of endosymbiotic green algae from the order Chlorellales, candidate bacterial endosymbionts, and bacterial food. We separated the initial read libraries into putative taxonomically distinct bins. Bins containing *Pb* signal were combined to form an initial assembly of 333 contigs totaling 29.95 Mbp, with an N50 of >130 kbp (where 90.64% of the genome is contained in contigs >50 kbp). As expected for *Paramecia* species (Aury et al. 2006; McGrath et al. 2014), this assembly had a low GC content of 27.22% (Fig. 1a).

The assembly profile is consistent with the short numerous chromosome structures typical of *Paramecia* macronuclei (Aury et al. 2006) and shows similar assembly statistics to previously sequenced *Pb* strains across most metrics (Fig. 1b and Table 1). Coverage of the genome assembly with the recovered PacBio HiFi-style reads (mean: 168.31×, SD: 85.33×), remaining PacBio CLR reads (mean: 134.38×, SD: 76.15×), and Illumina NovaSeq reads (mean: 4,730.97×, SD: 2,542.16×) was highly variable. Genome size estimation using GenomeScope 2.0 (Ranallo-Benavidez et al. 2020) and the Illumina NovaSeq data suggests a length between 30 and 34 Mbp, which is comparable to some previously sequenced *Pb* strains (Dd1 = 26.8 Mbp and 110224 = 29.2 Mbp) and *Paramecium caudatum* (*Pc*) with 30.5 Mbp and is only slightly larger than the total reported size of our 186b assembly of 29.9 Mbp (Table 1). BUSCO (Manni et al. 2021a, 2021b) analyses indicate a completion score of 58.4% using the Eukaryota ODB10 database, 57.8% using the Ciliate ODB12 database, and 88.3% using the Alveolate ODB10 database (using MetaEuk and "codon table 6" as the gene prediction software within BUSCO). The program OMArk (Nevers et al. 2022, 2024) (alignment free proteome

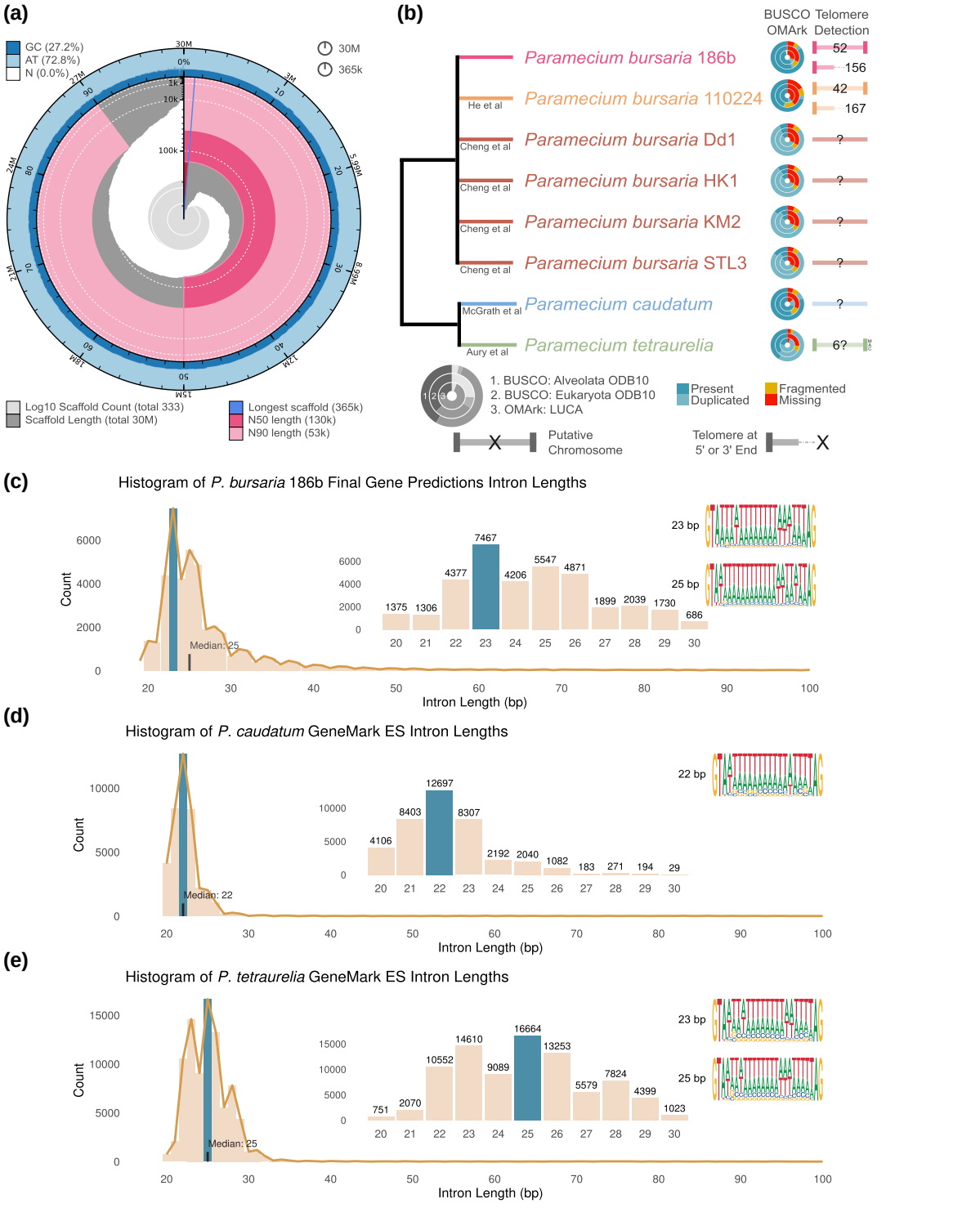

Fig. 1. Summary of genome assembly statistics and comparison with other *Paramecia* genomes. a) A snail diagram from Blobtools demonstrates the summary assembly data for *Pb* 186b, including GC content, overall size (Mbp), number and length of the longest scaffold(s), and including the N50/N90 statistics. Circumferential (30 M) and radial scales (365 k) are shown, top right. b) Summary cladogram showing a basic tree topology for three *Paramecia* species. Genome source data includes Aury et al. (2006), McGrath et al. (2014), He et al. (2019), and Cheng et al. (2020). Species nodes are labeled with BUSCO and OMArk assembly completion statistics (in concentric rings) and evidence of telomere structures is notated when recovered; see key for further details. Question marks denote no candidate telomeric sequences have been found and/or telomeres are unconfirmed in the original assembly source and associated publication. Intron size distribution and dominant intron sequences illustrated using gglogo (Schneider and Stephens et al. 1990) for *Pb* 186b (c), *Pc* (d), and *Pt* (e).

Table 1 Comparison of genome assembly data

	Leonard et al.	He et al.	Cheng et al.	Cheng et al.	Cheng et al.	Cheng et al.	McGrath et al.	Aury et al.
Assembly	*P.bur* 186b	*P.bur* 110224	*P.bur* Dd1	*P.bur* HK1	*P.bur* KM2	*P.bur* STL3	*P.cau*	*P.tet* d4-2
Database retrieval	NCBI: PRJNA659045	ParameciumDB	ParameciumDB	ParameciumDB	ParameciumDB	ParameciumDB	ParameciumDB	Ensembl
QUAST Analyses								
# of contigs	333	405	1,030	1,030	1,030	1,030	1,202	697
Total length (bp)	29,951,925	29,155,737	53,638,011	53,640,944	53,645,614	53,639,742	30,525,943	72,094,543
GC (%)	27.21	28.75	28.79	28.80	28.80	28.84	28.20	28.05
N50	130,222	96,293	98,752	98,603	98,773	98,624	313,711	413,286
L50	84	111	204	204	204	204	36	64
# of N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	2,168.23	799.45
Intron analyses								
Total	46,491	31,983	76,750	74,850	74,774	73,444	43,420	90,574
Genes with introns	17,418	11,406	26,665	26,039	26,099	25,757	15,178	31,675
Introns per gene	2.6	2.8	2.87	2.87	2.86	2.85	2.86	2.8 (2.3)
Mean intron length (bp)	28	27.60	27.50	27.39	27.37	27.42	23.29	25.13
BUSCO Alveolate (*n*: 171)								
Complete [single, duplicate]	88.3% [84.2%, 4.1%]	74.8% [71.3%, 3.5%]	86.0% [8.2%, 77.8%]	87.1% [9.9%, 77.2%]	87.1% [14.0%, 73.1%]	87.7% [15.2%, 72.5%]	87.7% [80.7%, 7.0%]	94.1% [41.5%, 52.6%]
Fragmented	4.1%	8.2%	7.0%	5.3%	5.8%	5.8%	6.4%	2.3%
Missing	7.6%	17.0%	7.0%	7.6%	7.1%	6.5%	5.9%	3.6%
OMArk Completeness								
Oligohymenophorea (*n*: 2612)								
Complete [single, duplicate]	69.4% [58.38%, 11.03%]	67.8% [15.4%, 52.3%]	68.2% [15.9%, 52.3%]	67.9% [15.6%, 52.2%]	67.8% [15.6%, 52.2%]	61% [52.6%, 8.4%]	86.1% [70.3%, 15.8%]	94% [23.8%, 70.2%]
Missing	30.59%	32.2%	31.7%	32.0%	32.1%	38.8%	13.8%	5.9%
OMArk Placement								
Consistent	64.34%	75.9%	73.6%	73.8%	73.6%	73.7%	81.0%	84.0%
Inconsistent	24.43%	15.6%	16.7%	16.5%	16.6%	16.5%	11.6%	9.3%
Contamination	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Unknown	11.2%	8.5%	9.7%	9.7%	9.8%	9.8%	7.4%	6.8%
GeneMark ES-only Predictions								
# of genes	21,279	11,529	26,289	26,314	26,215	26,027	15,158 (18,509)	33,526 (39,642)
# of introns	46,491	45,585	68,514	68,688	68,994	69,027	43,127	92,294
GenomeScope 2.0 genome size estimate (Mbp)	30–34	(29.2)	(26.8)	…	…	…	(30.5)	(87)

Numbers in parentheses are derived from genome portals or initial publication; all other numbers are from our reanalysis of these genomes and are generated by QUAST (Gurevich et al. 2013; Mikheenko et al. 2016) for comparison.

completeness assessment using the OMA orthology database) gave a completeness score of 69.4% using the LUCA dataset (with Oligohymenophorea identified as the best taxonomic affiliation).

Ciliate macronuclear genomes are often composed of numerous short chromosomes (Le Mouël et al. 2003; Eisen et al. 2006; Cheng et al. 2020). To further investigate chromosome structure in the *Pb* assemblies, we searched for identifiable telomere-like sequence tracks using both BLASTn-short and Tapestry searches (Davey et al. 2020, 2021) using the previously identified ciliate telomere sequence (Eisen et al. 2006) of "5′-CCCCAACCCCAA-3′" and its reverse complement "TTGGGGTTGGGG". To test for other repeat motifs which could represent variant telomere structures, we used TelFinder (Sun et al. 2023) which searches for repeat motifs via *k*-mer analysis at the ends of scaffolds. While several other putative repeat motifs were found, none were represented more than once across the scaffolds (at either end) and so they are unlikely to represent any varied, missed, or alternative telomere-like nucleotide sequence.

In total, we identified 208 (63.4%) contigs with telomere-like sequence motifs; of which, 52 (∼15%) contained telomere-like sequence motifs at both ends of the contig and 156 that contained a single telomere (∼47%). These numbers suggest that the nuclear genome sampled contained ∼130 (156/2 + 52) chromosomes in 333 fragments. The same analyses were conducted for all five additional publicly available *Pb* genome assemblies (He et al. 2019; Cheng et al. 2020), along with *Pc* (McGrath et al. 2014) and *Paramecium tetraurelia* (*Pt*) (Aury et al. 2006) genome sequence datasets, identifying additional "complete" chromosome-like structures in only the *Pb* 110224 and *Pt* assemblies (Fig. 1b). Considering only contigs with putative telomeres on both ends, these data demonstrate mean chromosome lengths of 89,973 bp ($n = 52$, median = 86,647 bp) for *Pb* 186b, 88,121 bp ($n = 42$, median = 79,007 bp) for *Pb* 110224, and 221,083 bp ($n = 6$, median = 284,010) for *Pt* (Fig. 1b). These data show some variability in chromosome size but are consistent with *Pb* 186b possessing small macronuclear chromosomes (Cheng et al. 2020), although these data could be affected by biased recovery leading to higher assembly completion rates among the shorter chromosomes. Indeed, if we consider the estimation of 130 chromosomes and the recovered genome size of 29.95 Mb as accurate, these data suggest a mean chromosome size of 0.23 Mb.

To further explore genome completion and to facilitate gene prediction and annotation, we generated a PacBio Iso-Seq transcriptome library. Gene predictions were completed using GeneMark-ES v7.4 (Lomsadze 2005) in ET mode, using the "–gcode 6" option (to account for alternative stop codon usage of ciliates), incorporating the Iso-Seq data—mapping intron locations to the genome assembly. This produced 21,279 putative genes, with a mean size of 1,352 bp (median of 965 bp), a mean intron size of 28 bp (median of 25 bp), and a mean exon size of 405 bp (median of 214 bp). The PacBio Iso-Seq refinement and clustering workflow (PacBio 2025a) includes a step for mapping the final reads to the genomic scaffolds. This identified 44,631 *Pb* candidate transcripts with 99.9% mapping directly to the genome assembly. This mapping result is consistent with a high level of completion for the macronuclear genome assembly.

Pb Intron Repertoire

Paramecia have been shown to possess a high number of short introns (Russell et al. 1994). This feature, combined with the reduced repertoire of stop codons within the ciliate nonstandard genetic code (Prescott 1994), means that identifying accurate open reading frames and, therefore, gene models is a significant challenge. Candidate introns were recovered by extracting putative intron sequences from the gene predictions (see Materials and Methods). These data demonstrate 46,491 introns, with 3,861 genes possessing no introns, 17,418 genes possessing one or more intron, and 11,475 genes with two or more introns. This is an average of 2.6 introns per gene, somewhat similar to other *Paramecia* when processed using the same bioinformatic pipeline (Table 1). These results also suggest a higher number of introns per gene for *Pt* (2.8 introns per gene using the same pipeline) than previously reported ((Russell et al. 1994; Saudemont et al. 2017)—2.3 introns per gene), but among a smaller number of gene models identified using the gene annotation pipeline applied here (Table 1). However, we note that intron predictions are sensitive to changes in gene annotation models, and the increase in intron recovery identified here may be due to changing sensitivity, different combinations of the bioinformatic software, and the type of transcriptome data used. Previous work suggests that *Pt* possesses a large number of "cryptic introns" (Russell et al. 1994; Saudemont et al. 2017), which may lead to variant estimations of intron number using different bioinformatic methods.

These data also showed a dominant proportion of 23 bp introns (highlighted in blue in Fig. 1c) in *Pb* 186b and all other *Pb* genomes (Fig. S1). In contrast, *Pc* has a dominant intron size of 22 bp (Fig. 1d), while *Pt* has a dominant intron size of 25 bp (Aury et al. 2006) (Fig. 1e). The dominance of 23 bp introns in the *Pb* genomes is notable as it is the same size as the small interfering RNAs (siRNAs) used by the RNAi pathway in both *Pb* and *Pt* (Jenkins et al. 2021b) to facilitate gene knockdown (Galvani and Sperling 2002), and endogenous small RNAs (sRNAs) of this size have been shown to regulate expression of both cis and trans gene targets in *Pt* (Karunanithi et al. 2019). Given the overlap in size of the dominant intron population and sRNA population in *Pb* 186b (Jenkins et al. 2021a), we tested for the possibility that the *Pb* 23 bp introns are present in the *Pb* sRNA population. To do this, we searched the assembled sRNA sequencing data from *Pb* 186b (Jenkins et al. 2021a) with

BLAT (enabling -minScore = 15), identifying no evidence that the *Pb* 23 bp introns are detectable among the *Pb* 186b sRNA transcripts. This is consistent with similar analysis from *Pt* (Carradec et al. 2015; Davey et al. 2021), which demonstrated a low rate of recovery for sRNAs containing intron sequences.

To explore intron sequence variation, we calculated sequence logos revealing that the 23 bp introns of *Pb* are largely composed of a core AT rich region and bear GT-AG splice sites like other *Paramecia* introns (Fig. 1c to e and Fig. S1) (Jaillon et al. 2008).

Alternative Splicing in *Pb*

The results reported here provide evidence for 21,279 predicted genes with 44,631 distinct Iso-Seq transcripts. This is suggestive of a substantial level of alternative splicing (AS) and is consistent with previous suggestions for other *Paramecia* species (Saudemont et al. 2017). To further explore these data, we detected AS events by mapping the Iso-Seq full-length non-concatemer transcripts using the program HISAT2 (Kim et al. 2019) onto the gene models and then used IsoQuant (Prjibelski et al. 2023) for isoform classification. This approach detected 29,322 candidate alternatively spliced events mapping to 6,532 *Pb* genes, demonstrating that ~30% of the *Pb* 186b gene repertoire shows putative evidence of AS.

Next, we used the IsoQuant pipeline to identify AS corresponding to putative intron retention (IR) events. IsoQuant identifies three categories of IR: "incomplete_intron_retention_right" which are IR events at the 3′ end of the mRNA that led to partial or truncated splice variants (183 events identified); "incomplete_intron_retention_left" which are IR events at the 5′ end that led to partial or truncated splice variants (163 events); and standard "intron_retention" splice variants (723 events) where the intron is retained within the core of the mRNA sequence. The IsoQuant pipeline identifies an additional category called "fake_micro_intron_retention", which includes annotated introns <50 bp in length which are described in the IsoQuant manual as "artifacts", i.e. "short annotated introns that are often missed or excluded by the aligners" (bioinformatic software is usually not designed with non-model organisms in mind, causing such analysis complications for other divergent taxa). Nonetheless, as demonstrated *Pb* possesses short introns, we checked all 4,912 of these events (~10% of the candidate AS events) across the alignments and found that they represent true IR events in all cases. These IR events correspond to 2,590 genes or ~12% of the gene repertoire and 8% of the AS events identified.

Conclusion

Here, we report a highly complete host *Pb* 186b macronuclear genome assembly along with replete transcriptome sequencing data using both Illumina and PacBio sequencing methods. This work complements our previous publications which include sRNA sequence datasets (Jenkins et al. 2021a) and a description of RNAi methodology for targeted knockdown of host-encoded genes (Jenkins et al. 2021b). These data are provided to aid genome analysis and to enable the design of forward and reverse genetic experiments for the 186b strain of *Pb*.

Materials and Methods

Pb 186b Culture Preparation and DNA/RNA Extraction

Cultures of *Pb* 186b (CCAP 1660/18) were grown in New Cereal Leaf—Prescott Liquid media (NCL). NCL media was prepared by adding 4.3 mgL^{-1} $CaCl_2.2H_2O$, 1.6 mgL^{-1} KCl, 5.1 mgL^{-1} K_2HPO_4, 2.8 mgL^{-1} $MgSO_4.7H_2O$ to deionized water. A 1 gL^{-1} wheat bran was added, and the solution boiled for 5 min. Once cooled, media was filtered once through Whatman Grade 1 filter paper and then through Whatman GF/C glass microfiber filter paper. Filtered NCL media was autoclaved at 121 °C for 30 min to sterilize prior to use.

NCL medium was bacterized with *Klebsiella pneumoniae* SMC and supplemented with 0.8 mgL^{-1} β-sitosterol prior to propagation. *Pb* cells were subcultured 1:9 into fresh bacterized NCL media every two months and maintained between 20 °C and 23 °C with a light-dark (LD) cycle of 12:12 h.

Pb cells for genomic DNA extractions were concentrated through centrifugation of 10 × 150 mL cultures (10 min at 800×*g*) followed by removal of ~80% of the supernatant. Concentrated cultures were then filtered using 15 µm PluriStrainers® and were washed repeatedly with sterile Milli-Q in order to reduce bacterial contamination. Genomic DNA was extracted from pooled cultures using the Qiagen DNeasy Blood and Tissue kit, and then the DNA was purified using a Qiagen DNeasy Power Cleanup kit. Long-read sequencing was performed using the SMRT Link software version 10.2.0.133434 on a Sequel IIe Pacific Biosciences (PacBio) device using 6× SMRT Cells following size selection (>3 kb) with AMPure PB Beads. The "Run Design Application" was set to CLR with default settings.

The resulting PacBio long-read sequence data comprised of six libraries: (i) Pb1_A05: 5,015,777 reads, 22,472,249,345 bp; (ii) Pb1_A08: 4,287,200 reads, 18,071,618,742 bp; (iii) Pb2_A01: 3,985,316 reads, 20,238,646,378 bp; (iv) Pb2_G10: 3,983,413 reads, 18,192,570,661 bp; (v) Pb3_A01: 3,019,403 reads, 13,420,246,768 bp; and (vi) Pb3_B01: 3,258,680 reads, 13,820,315,188 bp. In total, this includes 23,549,789 reads and 106,215,647,082 bp of sequence.

For Iso-Seq, ~750,000 *Pb* cells in stationary phase were harvested. For NovaSeq, six replicate *Pb* cultures were fed for 6 d with HT115 *E. coli* expressing non-hit, "scramble"

dsRNA (Jenkins et al. 2021a, 2021b) and ~500,000 *Pb* cells were harvested. For Iso-Seq and Illumina NovaSeq, *Pb* cells were collected on an 11 μm filter, washed with NCL, and rinsed into 1 mL of TRIzol reagent. RNA was extracted using the Zymoprep RNA extraction kit and stored in nuclease-free water at −20 °C. For RNA-Seq experiments, *Pb* samples were harvested at different points within the LD cycle (10.5 h into dark cycle "#1"; 6 h into light cycle "#3"; 1.5 h into dark cycle "#5"), as described previously (Jenkins et al. 2021a).

Short-Read DNA Sequencing Using an Illumina NovaSeq

DNA was processed using the Illumina NovaSeq 6000 v1.5 workflow (Illumina) with polyA selection. A 150 bp paired-end library was prepared and resulted in 1,184,917,880 reads total consisting of 177,737,682,000 bp.

RNA-Seq Sequencing Using Illumina

RNA from a LD cycle time-course (11 samples: #1A-C,E; #3A,C-E; #5B-D) was prepared for sequencing as described previously (Jenkins et al. 2021a).

Iso-Seq Sequencing Using Pacific Biosciences (PacBio)

RNA was processed using the Iso-Seq Express 2.0 workflow (PacBio), targeting transcripts up to 2 kb. Libraries were cleaned using the Express TPK 2.0 (PacBio) and SMRTbell Enzyme Clean-up kit v1 (PacBio) and prepared using the Sequel II Binding Kit 2.2 (PacBio), Sequencing Primer v5 (PacBio), and Sequel II Sequencing Plate v2.0 (PacBio). Sequencing was performed using SMRT link software (10.2.0.133434) on a Sequel IIe machine (PacBio). The "Run Design Application" was set to CCS (circular consensus sequencing) with default settings. The resulting PacBio HiFi long-read library consisted of 3,664,630 reads, with total length 8,734,233,994 bp, average length 2,383, and longest length 15,915.

Long-Read Genome Assembly

The BAM files from the Exeter Sequencing Service were converted to FASTQ using SAMTOOLS v1.15 (Li et al. 2009) and then concatenated together. The program LRBinner v2021-06-22 (Wickramarachchi and Lin 2022) was used to bin the reads using composition and coverage information via a variational auto-encoder. This resulted in 22 bins. These bins were then individually assembled using Flye (Kolmogorov et al. 2019) with standard settings. Following these preliminary assemblies, the BUSCO (Manni et al. 2021a, 2021b) tool in "auto-lineage" mode was used to assess each assembly bin for its basic taxonomic profile. This identified three bins (0, 1, and 2) as having strong alveolate signal, three other bins (3, 5, and 7) as having strong bacterial signal (these were not included in any assembly presented here), and all other bins as unclassified (also not included). Binning and subsequent classifications are not exact, and so some of the reads (and subsequent contigs) included may not represent the major taxonomic identity of the bin (i.e. the bins are not 100% clean and therefore may contain some cross "contamination"). The raw reads from the bins with strong alveolate signal (which made up the bulk of the sequencing libraries) were then further split into two sets. Although the sequencing performed in this experiment was PacBio CLR, we found that the bioinformatic part of the PacBio CCS (HiFi) pipeline could still be used to produce a set of high quality *HiFi-style* reads. All other reads, labeled as "failed" from this process, are in fact the remaining CLR reads. After splitting the initial reads into these two sets, we then used Flye again, but in a two-step hybrid assembly protocol utilizing both the HiFi-like and CLR reads. The resulting assembly was subsequently further polished using Pilon (Walker et al. 2014) and the Illumina NovaSeq reads (adapter trimmed using FastP (Chen et al. 2018)), making sure to trim any poly-G tails to account for the 2-color chemistry of the NovaSeq for one round. Finally, basic cleaning of the resulting contigs (removal of any duplicates) and repeat masking was completed with "funannotate clean & sort" (Palmer and Stajich 2020). Repeat masking was completed using both Repeat Modeler 2 (Flynn et al. 2020) and Repeat Masker v4.1.4 (Smit et al. 2025). This resulted in an assembly with 333 contigs, the largest being 364,712 bp, with an N50 of 130,222 bp.

Mitochondrion Assembly

The mitochondrial genomes of *P. aurelia* complex (NC_001324), *Pc* (NC_014262), and *P. gigas* and *T. thermophila* (NC_003029) were retrieved from NCBI using MitoHIFI (Uliano-Silva et al. 2023). *Pb* 110224's mitochondria were extracted from the main assembly (as contig: GWHAAFB00000001 (He et al. 2019). The aforementioned PacBio HiFi sequences were then mapped against these *mt* genomes using minimap2 (Li 2018, 2021); mapped reads were then extracted to fastq files and deduplicated using SeqKit2 (Shen et al. 2024). The resulting FASTQ was then assembled with Flye (Kolmogorov et al. 2019) producing one 56,624 bp contig.

Iso-Seq Assembly

The PacBio software "lima" (PacBio 2025b) (to remove barcodes/primers) and the "isoseq3 refine" and "isoseq3 cluster" (PacBio 2025a) (to remove polyA tails and cluster *de novo* isoforms, respectively) were used to prepare the data from the raw reads (reads: 3,664,630, total length: 8,734,233,994 bp, mean length: 2,383 bp, longest: 15,915 bp). The resulting 97,280 full-length non-concatemer transcripts (total length: 231,414,816 bp) were mapped to the genome assembly using "isoseq3 align" and then

isoforms were collapsed with "isoseq collapse". This produced 44,631 full-length transcripts. The final set of transcripts was then translated to amino acids using the "TransDecoder" (Haas 2025) pipeline with the Ciliate stop codon usage settings, producing 44,518 peptide sequences. The final set of transcripts was subjected to BUSCO (Manni et al. 2021a, 2021b) analysis and returned Eukaryota ODB10: complete:55.7% [single: 28.6%, duplicated: 27.1%], fragmented: 3.5%, missing: 40.8%, *n*: 255 and Alveolata ODB10: complete: 87.1% [single: 36.8%, duplicated: 50.3%], fragmented: 0.6%, missing: 12.3%, *n*: 171. This suggested that we had good coverage of the transcriptome. To further assess the coverage of the Iso-Seq transcriptome, we used pBLAT (Wang and Kong 2019) to search the transcript CDS against the genome, resulting in 44,111 matches at ≥97% identity suggesting 99% coverage (dropping to ≥87% ID returns at 100% coverage). Furthermore, 42,422 Iso-Seq transcripts map directly to 12,785 *Pb* mRNAs giving us a transcriptome representation of 60% of the candidate genes identified on the *Pb* genome.

Genome Annotation

Annotation of the cleaned genome assembly was conducted with GeneMark-ES (Ter-Hovhannisyan et al. 2008; Brůna et al. 2024) using the Ciliate genetic code table in "ET" mode and using introns identified from the mapping of the Iso-Seq transcripts to the genome assembly. Other settings used in this analysis included; constraining intergenic spaces to 100 bp minimum, max introns to 100 bp, 'intron_DUR min' to 10 bp, and the minimum gene size to 100 bp. Functional annotation was provided using multiple databases from InterProScan (PFAM` (Paysan Lafosse et al. 2024)), EGGNOG (Huerta-Cepas et al. 2016), BUSCO (Manni et al. 2021a, 2021b), Phobius (Madeira et al. 2019), and antiSMASH (Blin et al. 2021)) and integrated by the FUNANNOTATE (Palmer and Stajich 2020) pipeline in the "other" mode. Commands for the majority of processes in the methods can be found in the GitHub repository (https://github.com/guyleonard/paramecium).

Mitochondrion Genome Annotation

The tool Oatk (Zhou et al. 2024) was used to annotate the single contig representing a candidate mitochondrion assembly. Firstly, an OatkDB was built using the NCBI *Ciliophora* taxonomy ID "5878" in mitochondrion mode and setting the genetic code table to "4". This allowed the most coverage of genes from previously submitted mitochondrial genomes to be searched using HMM profiles generated by OatkDB. Secondly, the program Oatk takes the genome graph from the assembly along with HMM models created in the previous step and scans for matching candidate genes. Contig annotation revealed a ~11.4 kbp inverted repeat containing the genes rpl6, ymf64, atp9, nad1, rpl16, ymf65, yejR, nad3, and nad9 and confirms that the current mitochondrial assembly is not circular. Other mitochondrial genes located included cox I and cox II; cob; nad 1, 4, 5, 6, 7, and 10; rpl 2 and 14, rps 3, 12, 13, 14, and 19; and rrnS and rrnL, showing similar patterns to other *Paramecium* mitochondrial genomes (Johri et al. 2019).

Intron Identification and Mapping

Introns were extracted from the final set of gene predictions, using the script "agat_sp_add_introns.pl" (Dainat et al. 2023) which adds intron boundaries to the General Feature Format file (GFF) based on the predicted gene structures (exons are present in the GFF but introns are not commonly annotated directly). Their sequences were subsequently extracted using "bedtools getfasta" from the nuclear genome. This was repeated for the other five *Pb macronuclear* genomes (He et al. 2019; Cheng et al. 2020), *Pc* (McGrath et al. 2014), and *Pt* (Aury et al. 2006). All genomic data and gene prediction GFF files were downloaded from ParameciumDB (Arnaiz et al. 2019). These were then tallied in R (using phytools (Revell 2012), stringr (Wickham 2023), dplyr (Wickham 2025), Biostrings (Pagès et al. 2025), and ggplot2 (Wickham 2009)) (see GitHub for code).

AS Identification

The program IsoQuant (Prjibelski et al. 2023) was used with a copy of the final gene predictions in GFF format, the nuclear genomic scaffolds, and the full-length non-concatemer Iso-Seq transcripts mapped to the genome (using HISAT2 (Kim et al. 2019)) to produce a transcript table of all putative alternate splicing events and their genomic location paired with coverage data and the type of AS event identified.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Karen Moore and the University of Exeter Sequence Service for their support with the various sequencing projects. We thank Éric Meyer, Institut de Biologie de l'Ecole Normale Supérieure, Paris, for advice.

Funding

This work was primarily supported by an European Molecular Biology Organization Young Investigator Programme award and a Royal Society, University

Research Fellowship (UF130382) and latterly by an European Research Council, Consolidator Grant (CELL-in-CELL: 819507) to T.A.R.

Data Availability

All sequence reads have been deposited in NCBI GenBank with the BioProject identifier PRJNA659045 for PacBio: SRR12511009, SRR12511010, and SRR12511011; Illumina NovaSeq: SRR12511019; and PacBio Iso-Seq: SRR25546588. The RNA-Seq transcriptome data is available from BioProject PRJNA633103 as condition 1 (SRR11796780, SRR11796779, SRR11796768, SRR11796757), condition 3 (SRR11796746, SRR11796735, SRR11796724, SRR11796713), and condition 5 (SRR11796704, SRR11796703, SRR11796778). The *Pb* 186b genomic assembly is available at NCBI (JBOZOD010000000). The complete mitochondrion sequence can be accessed at PX289116. The genome annotations and other data (including key elements of code and commands) can also be accessed at https://github.com/guyleonard/paramecium and from Zenodo DOI: 10.5281/zenodo.13240530.

Literature Cited

Adl SM, et al. Revisions to the classification, nomenclature, and diversity of eukaryotes. J Eukaryot Microbiol. 2019:66:4–119. https://doi.org/10.1111/jeu.12691.

Archibald JM. The puzzle of plastid evolution. Curr Biol. 2009:19:R81–R88. https://doi.org/10.1016/j.cub.2008.11.067.

Arnaiz O, Meyer E, Sperling L. ParameciumDB 2019: integrating genomic data across the genus for functional and evolutionary biology. Nucleic Acids Res. 2019:48:D599–D605. https://doi.org/10.1093/nar/gkz948.

Aury J-M, et al. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature. 2006:444:171–178. https://doi.org/10.1038/nature05230.

Blin K, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res. 2021:49:W29–W35. https://doi.org/10.1093/nar/gkab335.

Bomford R. Infection of alga-free *Paramecium bursaria* with strains of *Chlorella, Scenedesmus*, and a yeast. J Protozool. 1965:12:221–224. https://doi.org/10.1111/j.1550-7408.1965.tb01840.x.

Bonen L, Cunningham RS, Gray MW, Doolittle WF. Wheat embryo mitochondrial 18S ribosomal RNA: evidence for its prokaryotic nature. Nucleic Acids Res. 1977:4:663–671. https://doi.org/10.1093/nar/4.3.663.

Bonen L, Doolittle WF. On the prokaryotic nature of red algal chloroplasts. Proc Natl Acad Sci U S A. 1975:72:2310–2314. https://doi.org/10.1073/pnas.72.6.2310.

Brůna T, Lomsadze A, Borodovsky M. GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. Genome Res. 2024:34:757–768. https://doi.org/10.1101/gr.278373.123.

Carradec Q, et al. Primary and secondary siRNA synthesis triggered by RNAs from food bacteria in the ciliate *Paramecium tetraurelia*. Nucleic Acids Res. 2015:43:1818–1833. https://doi.org/10.1093/nar/gku1331.

Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018:34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

Cheng Y-H, et al. Genome plasticity in *Paramecium bursaria* revealed by population genomics. BMC Biol. 2020:18:180. https://doi.org/10.1186/s12915-020-00912-2.

Curtis BA, et al. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature. 2012:492:59–65. https://doi.org/10.1038/nature11681.

Dainat J, et al. NBISweden/AGAT: AGAT v1.5.1 (v1.5.1). Zenodo. https://doi.org/10.5281/zenodo.16317950.

Davey JW et al. Chromosomal assembly of the nuclear genome of the endosymbiont-bearing trypanosomatid *Angomonas deanei*. G3 (Bethesda). 2021:11:jkaa018. https://doi.org/10.1093/g3journal/jkaa018.

Davey JW, Davis SJ, Mottram JC, Ashton PD. Tapestry: validate and edit small eukaryotic genome assemblies with long reads. bioRxiv. 2020. https://doi.org/10.1101/2020.04.24.059402.

Eisen JA, et al. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol. 2006:4:e286. https://doi.org/10.1371/journal.pbio.0040286.

Flynn JM, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020:117:9451–9457. https://doi.org/10.1073/pnas.1921046117.

Galvani A, Sperling L. RNA interference by feeding in paramecium. Trends Genet. 2002:18:11–12. https://doi.org/10.1016/s0168-9525(01)02548-3.

Greczek-Stachura M, et al. Identification of *Paramecium bursaria* syngens through molecular markers—comparative analysis of three loci in the nuclear and mitochondrial DNA. Protist. 2012:163:671–685. https://doi.org/10.1016/j.protis.2011.10.009.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013:29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.

Haas B. 2025. TransDecoder/TransDecoder: TransDecoder source. https://github.com/TransDecoder/TransDecoder (Accessed 5 August 2024).

He M, et al. Genetic basis for the establishment of endosymbiosis in *Paramecium*. ISME J. 2019:13:1360–1369. https://doi.org/10.1038/s41396-018-0341-4.

Huerta-Cepas J, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016:44:D286–D293. https://doi.org/10.1093/nar/gkv1248.

Jaillon O, et al. Translational control of intron splicing in eukaryotes. Nature. 2008:451:359–362. https://doi.org/10.1038/nature06495.

Jenkins BH, et al. Characterization of the RNA-interference pathway as a tool for reverse genetic analysis in the nascent phototrophic endosymbiosis, *Paramecium bursaria*. R Soc Open Sci. 2021a:8:210140. https://doi.org/10.1098/rsos.210140.

Jenkins BH, et al. Emergent RNA–RNA interactions can promote stability in a facultative phototrophic endosymbiosis. Proc Natl Acad Sci U S A. 2021b:118:e2108874118. https://doi.org/10.1073/pnas.2108874118.

Johri P, Marinov GK, Doak TG, Lynch M. Population genetics of paramecium mitochondrial genomes: recombination, mutation spectrum, and efficacy of selection. Genome Biol Evol. 2019:11:1398–1416. https://doi.org/10.1093/gbe/evz081.

Karakashian SJ, Karakashian MW. Evolution and symbiosis in the genus chlorella and related algae. Evolution. 1965:19:368. https://doi.org/10.2307/2406447.

Karunanithi S, et al. Exogenous RNAi mechanisms contribute to transcriptome adaptation by phased siRNA clusters in *Paramecium*. Nucleic Acids Res. 2019:47:8036–8049. https://doi.org/10.1093/nar/gkz553.

Kato Y, Imamura N. Metabolic control between the symbiotic chlorella and the host paramecium. In: Fujishima M, editors. Endosymbionts

in Paramecium. Springer; 2009. p. 57–82. https://doi.org/10.1007/978-3-540-92677-1_3.
Keeling PJ. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. Annu Rev Plant Biol. 2013:64:583–607. https://doi.org/10.1146/annurev-arplant-050312-120144.
Kessler E, Kauer G, Rahat M. Excretion of sugars by *Chlorella* species capable and incapable of symbiosis with *Hydra viridis*. Bot Acta. 1991:104:58–63. https://doi.org/10.1111/j.1438-8677.1991.tb00194.x.
Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019:37:907–915. https://doi.org/10.1038/s41587-019-0201-4.
Kodama Y, Fujishima M. Synchronous induction of detachment and reattachment of symbiotic Chlorella spp. from the cell cortex of the host *Paramecium bursaria*. Protist. 2013:164:660–672. https://doi.org/10.1016/j.protis.2013.07.001.
Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019:37:540–546. https://doi.org/10.1038/s41587-019-0072-8.
Kwong WK, Del Campo J, Mathur V, Vermeij MJA, Keeling PJ. A widespread coral-infecting apicomplexan with chlorophyll biosynthesis genes. Nature. 2019:568:103–107. https://doi.org/10.1038/s41586-019-1072-z.
Le Mouël A, Butler A, Caron F, Meyer E. Developmentally regulated chromosome fragmentation linked to imprecise elimination of repeated sequences in paramecia. Eukaryot Cell. 2003:2:1076–1090. https://doi.org/10.1128/ec.2.5.1076-1090.2003.
Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009:25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018:34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 2021:37:4572–4574. https://doi.org/10.1093/bioinformatics/btab705.
Lomsadze A. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 2005:33:6494–6506. https://doi.org/10.1093/nar/gki937.
Lowe CD, Minter EJ, Cameron DD, Brockhurst MA. Shining a light on exploitative host control in a photosynthetic endosymbiosis. Curr Biol. 2016:26:207–211. https://doi.org/10.1016/j.cub.2015.11.052.
Madeira F, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019:47:W636–W641. https://doi.org/10.1093/nar/gkz268.
Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021a:38:4647–4654. https://doi.org/10.1093/molbev/msab199.
Manni M, Berkeley MR, Seppey M, Zdobnov EM. BUSCO: assessing genomic data quality and beyond. Curr Protoc. 2021b:1:e323. https://doi.org/10.1002/cpz1.323.
McCutcheon JP, Garber AI, Spencer N, Warren JM. How do bacterial endosymbionts work with so few genes? PLOS Biol. 2024:22:e3002577. https://doi.org/10.1371/journal.pbio.3002577.
McGrath CL, Gout J-F, Doak TG, Yanagi A, Lynch M. Insights into three whole-genome duplications gleaned from the *Paramecium caudatum* genome sequence. Genetics. 2014:197:1417–1428. https://doi.org/10.1534/genetics.114.163287.
Mikheenko A, Valin G, Prjibelski A, Saveliev V, Gurevich A. Icarus: visualizer for *de novo* assembly evaluation. Bioinformatics. 2016:32:3321–3323. https://doi.org/10.1093/bioinformatics/btw379.
Minter EJA, et al. Variation and asymmetry in host-symbiont dependence in a microbial symbiosis. BMC Evol Biol. 2018:18:108. https://doi.org/10.1186/s12862-018-1227-9.
Muscatine L, Karakashian SJ, Karakashian MW. Soluble extracellular products of algae symbiotic with a ciliate, a sponge and a mutant hydra. Comp Biochem Physiol. 1967:20:1–12. https://doi.org/10.1016/0010-406X(67)90720-7.
Nevers Y, et al. Multifaceted quality assessment of gene repertoire annotation with OMArk. bioRxiv. 2022. https://doi.org/10.1101/2022.11.25.517970.
Nevers Y, et al. Quality assessment of gene repertoire annotations with OMArk. Nat Biotechnol. 2025:43:124–133. https://doi.org/10.1038/s41587-024-02147-w.
PacBio. 2025a. Iso-Seq Home. Iso-Seq Docs. https://isoseq.how/ (Accessed 5 August 2024).
PacBio. 2025b. Lima Home. Lima Docs. https://lima.how/ (Accessed 5 August 2024).
Pagès H, Aboyoun P, Gentleman R, DebRoy S. 2025. Biostrings. Bioconductor. http://bioconductor.org/packages/Biostrings/ (Accessed August 5, 2024).
Palmer JM, Stajich J. 2020. Funannotate v1.8.1: Eukaryotic genome annotation. https://doi.org/10.5281/zenodo.4054262.
Paysan-Lafosse T, et al. The Pfam protein families database: embracing AI/ML. Nucleic Acids Res. 2024:53:D523–D534.
Prescott DM. The DNA of ciliated protozoa. Microbiol Rev. 1994:58:233–267. https://doi.org/10.1128/mr.58.2.233-267.1994.
Prjibelski AD, et al. Accurate isoform discovery with IsoQuant using long reads. Nat Biotechnol. 2023:41:915–918. https://doi.org/10.1038/s41587-022-01565-y.
Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020:11:1432. https://doi.org/10.1038/s41467-020-14998-3.
Reisser W. Naturally occurring and artificially established associations of ciliates and algae. Ann N Y Acad Sci. 1987:503:316–329. https://doi.org/10.1111/j.1749-6632.1987.tb40618.x.
Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 2012:3:217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x.
Russell CB, Fraga D, Hinrichsen RD. Extremely short 20–33 nucleotide introns are the standard length in *Paramecium tetraurelia*. Nucleic Acids Res. 1994:22:1221–1225. https://doi.org/10.1093/nar/22.7.1221.
Saudemont B, et al. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. Genome Biol. 2017:18:208. https://doi.org/10.1186/s13059-017-1344-6.
Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990:18:6097–100. https://github.com/heike/gglogo (Accessed 5 August 2024).
Shen W, Sipos B, Zhao L. SeqKit2: a Swiss army knife for sequence and alignment processing. iMeta. 2024:3:e191. https://doi.org/10.1002/imt2.191.
Siegel RW. Hereditary endosymbiosis in *Paramecium bursaria*. Exp Cell Res. 1960:19:239–252. https://doi.org/10.1016/0014-4827(60)90005-7.
Smit AFA, Hubley R, Green P. RepeatMasker frequently asked questions. 2025. https://www.repeatmasker.org/faq.html (Accessed June 16, 2025).
Sørensen MES, Wood AJ, Cameron DD, Brockhurst MA. 2021. Rapid compensatory evolution can rescue low fitness symbioses following partner switching. Curr Biol. 31:3721–3728.e4. doi: 10.1016/j.cub.2021.06.034.
Spanner C, Darienko T, Biehler T, Sonntag B, Pröschold T. Endosymbiotic green algae in *Paramecium bursaria*: a new isolation method and a

simple diagnostic PCR approach for the identification. Diversity (Basel). 2020:12:240. https://doi.org/10.3390/d12060240.
Spanner C, Darienko T, Filker S, Sonntag B, Pröschold T. Morphological diversity and molecular phylogeny of five *Paramecium bursaria* (Alveolata, Ciliophora, Oligohymenophorea) syngens and the identification of their green algal endosymbionts. Sci Rep. 2022:12:18089. https://doi.org/10.1038/s41598-022-22284-z.
Sun Q, Wang H, Tao S, Xi X. Large-scale detection of telomeric motif sequences in genomic data using TelFinder. Microbiol Spectr. 2023:11:e03928-22. https://doi.org/10.1128/spectrum.03928-22.
Takeda H, Sekiguchi T, Nunokawa S, Usuki I. Species-specificity of *Chlorella* for establishment of symbiotic association with *Paramecium bursaria*—does infectivity depend upon sugar components of the cell wall? Eur J Protistol. 1998:34:133–137. https://doi.org/10.1016/S0932-4739(98)80023-0.
Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. 2008:18:1979–1990. https://doi.org/10.1101/gr.081612.108.
Uliano-Silva M, et al. Mitohifi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. BMC Bioinformatics. 2023:24:288. https://doi.org/10.1186/s12859-023-05385-y.
Walker BJ, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014:9:e112963. https://doi.org/10.1371/journal.pone.0112963.
Wang M, Kong L. pblat: a multithread blat algorithm speeding up aligning sequences to genomes. BMC Bioinformatics. 2019:20:28. https://doi.org/10.1186/s12859-019-2597-8.
Wickham H. Ggplot2: elegant graphics for data analysis. Springer; 2009. https://doi.org/10.1007/978-0-387-98141-3.
Wickham H. 2023. stringr: Simple, Consistent Wrappers for Common String Operations. https://stringr.tidyverse.org/ (Accessed August 5, 2024).
Wickham H. dplyr: a grammar of data manipulation. 2025. https://dplyr.tidyverse.org/. (Accessed August 5, 2024).
Wickramarachchi A, Lin Y. Binning long reads in metagenomics datasets using composition and coverage information. Algorithms Mol Biol. 2022:17:14. https://doi.org/10.1186/s13015-022-00221-z.
Zhou C, et al. Oatk: a de novo assembly tool for complex plant organelle genomes. bioRxiv. 2024. https://doi.org/10.1101/2024.10.23.619857.

Associate editor: Michael Lynch