


RESEARCH ARTICLE

Open Access

# Genome plasticity in *Paramecium bursaria* revealed by population genomics



Yu-Hsuan Cheng<sup>1,2</sup>, Chien-Fu Jeff Liu<sup>2</sup>, Yen-Hsin Yu<sup>2</sup>, Yu-Ting Jhou<sup>2</sup>, Masahiro Fujishima<sup>4</sup>, Isheng Jason Tsai<sup>1,3</sup> and Jun-Yi Leu<sup>1,2\*</sup> 

## Abstract

**Background:** Ciliates are an ancient and diverse eukaryotic group found in various environments. A unique feature of ciliates is their nuclear dimorphism, by which two types of nuclei, the diploid germline micronucleus (MIC) and polyploidy somatic macronucleus (MAC), are present in the same cytoplasm and serve different functions. During each sexual cycle, ciliates develop a new macronucleus in which newly fused genomes are extensively rearranged to generate functional minichromosomes. Interestingly, each ciliate species seems to have its way of processing genomes, providing a diversity of resources for studying genome plasticity and its regulation. Here, we sequenced and analyzed the macronuclear genome of different strains of *Paramecium bursaria*, a highly divergent species of the genus *Paramecium* which can stably establish endosymbioses with green algae.

**Results:** We assembled a high-quality macronuclear genome of *P. bursaria* and further refined genome annotation by comparing population genomic data. We identified several species-specific expansions in protein families and gene lineages that are potentially associated with endosymbiosis. Moreover, we observed an intensive chromosome breakage pattern that occurred during or shortly after sexual reproduction and contributed to highly variable gene dosage throughout the genome. However, patterns of copy number variation were highly correlated among genetically divergent strains, suggesting that copy number is adjusted by some regulatory mechanisms or natural selection. Further analysis showed that genes with low copy number variation among populations tended to function in basic cellular pathways, whereas highly variable genes were enriched in environmental response pathways.

**Conclusions:** We report programmed DNA rearrangements in the *P. bursaria* macronuclear genome that allow cells to adjust gene copy number globally according to individual gene functions. Our results suggest that large-scale gene copy number variation may represent an ancient mechanism for cells to adapt to different environments.

**Keywords:** Ciliate, *Paramecium*, Copy number variation, Programmed DNA rearrangement, Comparative genomics, Minichromosomes

\* Correspondence: [jleu@imb.sinica.edu.tw](mailto:jleu@imb.sinica.edu.tw)

<sup>1</sup>Genome and Systems Biology Degree Program, Academia Sinica and National Taiwan University, Taipei 106, Taiwan

<sup>2</sup>Institute of Molecular Biology, Academia Sinica, 128 Sec. 2, Academia Road, Nankang, Taipei 115, Taiwan

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Copy number variation (CNV) resulting from segmental DNA duplications or deletions (usually  $\geq 50$  base pairs, bp) represents a crucial source of genetic variation contributing to phenotypic diversity [1, 2]. For decades, most studies of phenotypic diversity focused on single nucleotide polymorphisms (SNPs), whereas the effects of CNV remain understudied [3]. However, owing to improved sequencing techniques, the prevalence and importance of CNV in model organisms are gradually being revealed [4–8].

Most large-scale studies on CNV have been conducted on human populations, which show that de novo CNV frequently occurs in human genomes [9, 10] and accounts for 17.7% of altered gene expression among genes associated with CNV or SNPs [11]. Moreover, CNV has been linked to a broad range of genetic diseases and complex traits [10, 12]. More recently, the phenotypic outcomes of CNV have also been investigated in other organisms, including domestication traits in animals and plants [13], chemical or disease resistance in insects and plants [14, 15], and environmental adaptation in microorganisms [3, 16]. These studies provide examples of how organismal phenotypes are influenced by CNV. However, it remains unclear to what extent CNV can be tolerated in a typical genome and if organisms systematically utilize CNV to adjust their physiologies under different environments.

Ciliates are excellent models for studying tolerance to and regulation of CNV [17–21]. A unique feature of ciliates is their nuclear dimorphism, by which two types of nuclei are present in the same cytoplasm and serve different functions throughout the life cycle [22]. The micronucleus (MIC) contains the diploid and transcriptionally silent germline genome, whereas the macronucleus (MAC) harbors the polyploid and transcriptionally active somatic genome. During asexual reproduction, the MIC undergoes typical mitosis and the MAC divides by amitosis, a type of nuclear division that does not involve spindle fibers. Amitotic nuclear division does not precisely segregate duplicated chromosomes, probably allowing ciliates to have a highly plastic MAC genome. To date, the underlying mechanism of amitotic chromosome segregation has remained largely uncharacterized. In each round of sexual reproduction, only the genetic material in the MIC is passed on to the progeny. The existing MAC is degraded after fertilization and a new MAC develops from the endoduplicated zygote. During this developmental period, a series of large-scale programmed DNA rearrangement events occur including DNA amplification, chromosome breakage, and excision of internal eliminated sequences (IESs) and repeated sequences. These rearrangements result in the formation of minichromosomes that can further increase the flexibility of CNV regulation.

Various types of genome rearrangements have been identified during MAC development in different clades of ciliates [23–26]. For instance, in *Tetrahymena thermophila*, five pairs of chromosomes are broken at specific sites and telomeric repeats are added de novo [27–30], resulting in the formation of approximately 181 specific minichromosomes averaging 68 copies each [31–33]. In *Paramecium tetraurelia*, the chromosomes undergo precise elimination of IESs and imprecise removal of repeated elements and transposons, and they are highly amplified to around 800 copies [34]. After DNA elimination, the ends are either joined by non-homologous end joining (NHEJ) or telomeres are added de novo to form minichromosomes. For both *T. thermophila* and *P. tetraurelia*, minichromosomes are maintained at constant copy numbers after conjugation. However, it has been observed in *P. tetraurelia* that minichromosomes might be fragmented by DNA damage and become shorter during clonal aging [35]. In *Oxytricha trifallax*, after the IESs have been removed, its chromosomes are extensively fragmented and unscrambled to form thousands of nanochromosomes that typically carry only one to eight genes [36, 37]. Nanochromosome copy number can be further adjusted by means of a maternal RNA-mediated mechanism during MAC differentiation [20, 21, 38].

*Paramecium bursaria* is one of only two species in the genus *Paramecium* that harbor algal endosymbionts [39, 40]. Based on a phylogenetic tree constructed from *Paramecium* 18S rRNA sequences with *T. thermophila* as outgroup, *P. bursaria* is the most diverged species since the most common *Paramecium* ancestor [41], which may explain why *P. bursaria* cell physiology is so distinct from other *Paramecium* species. In the wild, most *P. bursaria* cells stably harbor several hundred algal cells in the cytoplasm [42]. Each algal cell is engulfed by a host-derived perialgal vacuole (PV) membrane that becomes attached to the host plasma membrane [43, 44]. Different species of algal endosymbionts have been identified in different strains of *P. bursaria* [45], suggesting that these strains may have evolved a preference for or compatibility with specific green algae. Thus, *P. bursaria* strains and their endosymbiotic algae represent a useful system for understanding how stable endosymbiosis is initiated and maintained.

To facilitate genomic analyses of *P. bursaria*, we sequenced and annotated the functional macronuclear genome of *P. bursaria*. Unlike *T. thermophila* in which individual minichromosomes are maintained in similar copy numbers [31, 32], we observed extensive CNV at both chromosome and gene levels. By comparing the genomes of different *P. bursaria* strains, we reveal that minichromosomes maintained in consistent copy numbers among strains are enriched with housekeeping genes, whereas those carrying environmental response

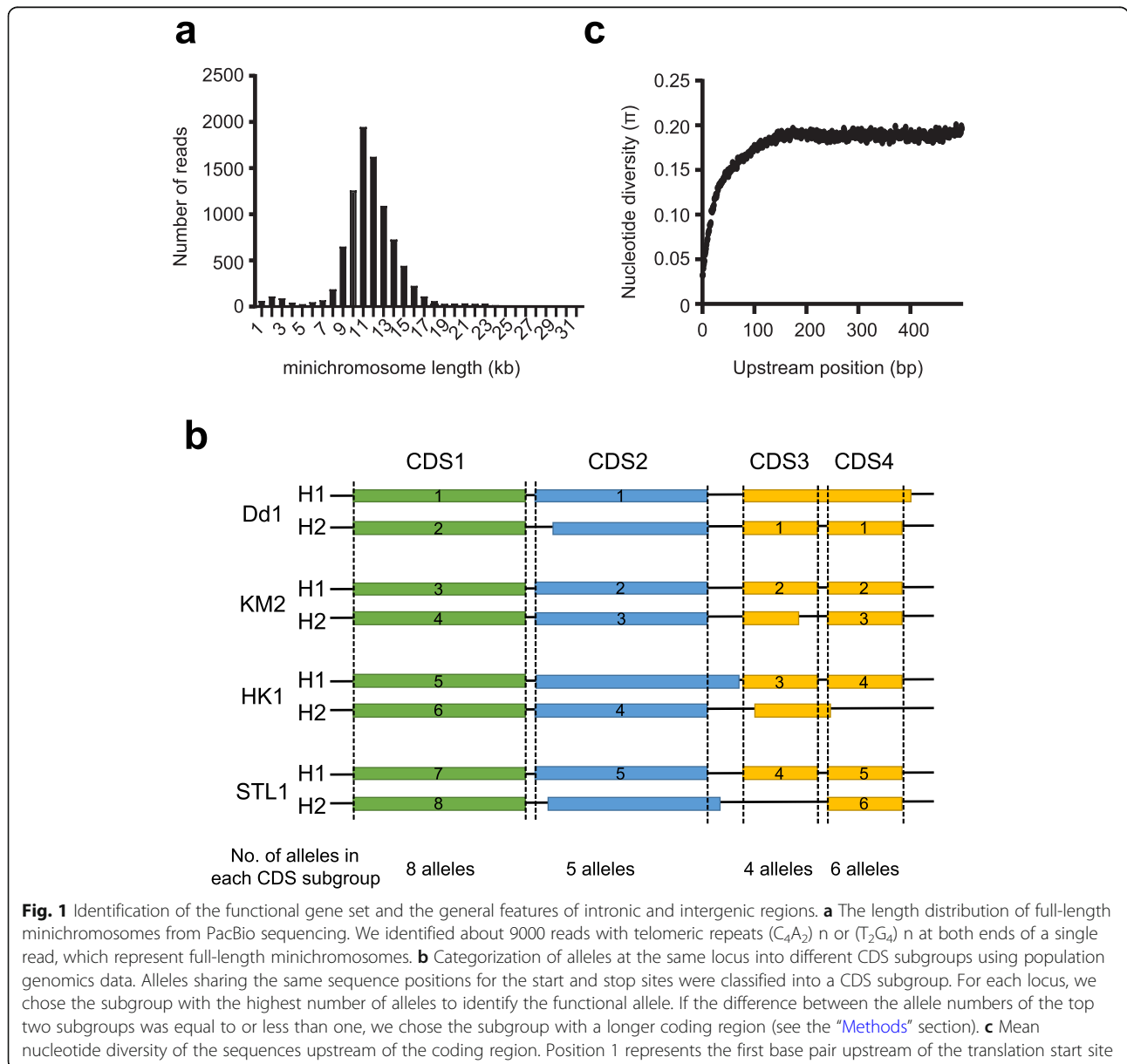
genes exhibit pronounced CNV among strains. Our results suggest that *P. bursaria* exhibits high plasticity and tolerance to gene copy number variation, which may facilitate cell adaptation to different environments.

**Results**

**A global view of the *Paramecium bursaria* macronuclear genome**

To assemble a macronuclear reference genome of *P. bursaria*, we isolated macronuclei from an aposymbiotic strain (i.e., cells without endosymbionts), Dd1, and sequenced the genomic DNA using both Illumina and PacBio sequencing platforms (Additional file 1: Fig. S1a). From the PacBio long-read data, we noticed that most

full-length minichromosomes (with telomeric repeats at both ends) were around 8 to 16 kb in length (Fig. 1a). Moreover, even the minichromosomes with similar gene contents had highly variable breakage sites. To reduce the complexity of the reference genome, we decided to use the overlapping regions between different minichromosomes and long reads to assemble artificial long contigs (see the “Methods” section). Among 597 contigs of the initial 30.1 Mb assembly, 129 contigs shared greater than 70% sequence identity to other contigs, perhaps due to highly divergent haplotypes. We removed redundant parts of these contigs, preserving only regions containing unique coding sequences (CDSs), and renamed them as regions of internal structural variation (ISVs;



**Fig. 1** Identification of the functional gene set and the general features of intronic and intergenic regions. **a** The length distribution of full-length minichromosomes from PacBio sequencing. We identified about 9000 reads with telomeric repeats ( $C_4A_2$ )<sub>n</sub> or ( $T_2G_4$ )<sub>n</sub> at both ends of a single read, which represent full-length minichromosomes. **b** Categorization of alleles at the same locus into different CDS subgroups using population genomics data. Alleles sharing the same sequence positions for the start and stop sites were classified into a CDS subgroup. For each locus, we chose the subgroup with the highest number of alleles to identify the functional allele. If the difference between the allele numbers of the top two subgroups was equal to or less than one, we chose the subgroup with a longer coding region (see the “Methods” section). **c** Mean nucleotide diversity of the sequences upstream of the coding region. Position 1 represents the first base pair upstream of the translation start site

Additional file 1: Fig. S1b). After manually examining the alignment, we further scaffolded 55 contigs by means of end sequence homology. The final assembly included 413 contigs and 102 ISVs totaling 26.8 Mb (Table 1). The GC content of the genome (28.8%) is similar to that reported for other *Paramecium* species [46, 47]. Combining the information from the size of our assembled MAC genome and previous microspectrophotometry data [52], we estimated that the average gene copy number is around 6000 in MACs.

An initial analysis revealed that the sequence diversity of the two haplotypes was high in our sequenced genome, which may result in inaccurate gene annotation due to misassignment of two alleles. We performed haplotype phasing to construct both haplotypes and then carried out gene annotation by combining de novo gene prediction and RNA sequencing (Additional file 1: Fig. S1a, Additional file 2: Table S1, and Methods). We found that sequence divergence between the two haplotypes was on average 2.5%, which is significantly higher than reported for several well-studied eukaryotes [53–55]. We annotated about 15,600 CDSs from each haplotype and ~40% (6273/15,601) either exhibited different start or/and stop sites between two alleles or were only present in one haplotype (Fig. 1b), suggesting that a large proportion of the proteome may be truncated and/or have lost function.

To construct a functional gene set, we further sequenced the macronuclear genomes of another three genetically divergent *P. bursaria* strains (KM2, HK1, and STL3). If we detected similar gene structures or alleles among these different strains, they would more likely

represent functional genes/alleles (Fig. 1b and the “Methods” section). The genomes of these three strains exhibited a sequence diversity ranging from 2.1 to 4.0% relative to the Dd1 reference genome (Additional file 2: Table S2), providing useful population data for annotating the functional genome. In total, 15,101 genes were classified into our functional gene set. To evaluate the completeness of our assembled genome, we examined it using the Core Eukaryotic Gene Mapping Approaches (CEGMA) database and identified 225 of 248 (91%) eukaryotic core genes with BLASTP [56]. Compared to the 220–230 core genes identified in the *Tetrahymena*, *Oxytricha*, and other *Paramecium* genomes [31, 36, 46, 47], this outcome indicates that our assembly is comparably complete.

Similar to other *Paramecium* species [46, 47, 57], the *P. bursaria* genome has extremely short intergenic regions and introns (Additional file 2: Table S3, Additional file 2: Fig. S1d and S1e), suggesting that the transcriptional regulation and splicing machineries of *P. bursaria* deviate considerably from non-ciliate organisms. Functional parts of noncoding regions are typically more conserved than nonfunctional sequences when different organismal populations are compared [58, 59]. We used population data to calculate the nucleotide diversity ( $\pi$ ) of the noncoding regions upstream of the translation start sites of individual genes. We found that sequence diversity gradually increased from the translation start site and reached a plateau after 150 bp (Fig. 1c). These data suggest that regulatory elements are mainly located within 50 bp upstream of the coding region, which in *P.*

**Table 1** Comparison of different ciliate macronuclear genomes

	<i>Paramecium bursaria</i>	<i>Paramecium caudatum</i> <sup>a</sup>	<i>Paramecium tetraurelia</i> <sup>b</sup>	<i>Tetrahymena thermophila</i> <sup>c</sup>	<i>Oxytricha trifallax</i> <sup>d</sup>
Genome size (Mb)	26.8	30.5	72.1	103	55.4
Sequencing platform	PacBio and Illumina	Illumina	Shotgun sequencing and Illumina	Shotgun sequencing and Illumina	PacBio and Illumina
No. of contigs	515 (413 + 102 ISV)	1202	697	1158	19,152
Size distribution of chromosomes observed using pulsed-field electrophoresis (Kb)	6–90 <sup>e</sup>	50–750 <sup>f</sup>	50–1000 <sup>g</sup>	21–1500 <sup>f</sup>	0.4–40 <sup>h</sup>
N50 (Kb)	100	312.9	413	520	3.5
Longest contig (Kb)	250	793	980	2216	66
Genomic GC content (%)	28.8	28.2	28.0	22.0	31.0
Gene number	15,101	18,673	40,460	26,996	18,400
Average gene length (bp)	1513	1458	1427	2400	1839

ISV internal structure variation region

<sup>a</sup>Taken from [46]

<sup>b</sup>Taken from [47, 48]

<sup>c</sup>Taken from [31, 49]

<sup>d</sup>Taken from [37]

<sup>e</sup>This study

<sup>f</sup>Taken from [50]

<sup>g</sup>Taken from [47]

<sup>h</sup>Taken from [51]

*bursaria* is more compact than in other well-studied eukaryotic organisms [59]. We also examined all the introns and discovered that the 5' (GT) and 3' (AG) splice sites remained conserved (Additional file 1: Fig. S2 [60]). Nonetheless, of the 92 splicing-related genes in *S. cerevisiae*, we only identified 66 orthologs in our assembled *P. bursaria* genome (Additional file 3: Table S4). The missing genes include some important components that are essential for yeast viability (*AAR2*, *CWC25*, *LSM8*, *PRP24*, *PRP39*, *PRP42*, *SNU114*, and *SPP2*) and that are shared between yeast and human cells.

### Protein family expansion and gene duplication in *Paramecium bursaria*

Among the known *Paramecium* species, *P. bursaria* is the most divergent lineage [41] and one of only two species that can stably establish endosymbiotic relationships with *Chlorella* spp. To gain insights into the specific physiology of *P. bursaria*, we compared the functional gene set of *P. bursaria* with those of another two sequenced *Paramecium* genomes, *P. caudatum*, and *P. tetraurelia*. We found that 75% (11,328/15,101) of *P. bursaria* genes have orthologs in these other *Paramecium* species (Fig. 2a). Among 3773 *P. bursaria*-specific genes, gene ontology (GO) terms related to phosphorylation signal transduction, drug transmembrane transport, and vesicle-mediated transport were significantly enriched (Additional file 2: Table S5). Genes related to these GO terms are also known to be expanded in *T. thermophila* relative to animals and yeast [31]. The observed enrichment suggests that some existing pathways may have been further modified in *P. bursaria* to innovate lineage-specific phenotypes.

To investigate if specific protein families are expanded in *P. bursaria*, we compared the gene numbers of all known protein families between *P. bursaria*, *P. caudatum*, and *P. tetraurelia*. Only one protein family was specifically enriched in the *P. bursaria* genome (Fig. 2b); the coatamer complexes coat membrane-bound vesicles and play an important role in vesicle transportation [62, 63]. Previous studies in *P. bursaria* have shown that during the establishment of endosymbiosis, green algae are always coated with a PV membrane and transported beneath the host cell cortex [64, 65]. Moreover, pathogens are known to hijack the coatamer-mediated pathway to enter and populate host cells [66]. Expansion of the coatamer family in *P. bursaria* probably represents a co-evolutionary event specific to algal endosymbiosis.

Gene duplication has also been suggested to contribute to specific adaptation during organism evolution [67, 68]. We further analyzed lineage-specific enrichment of gene duplication using OrthoFinder [61]. We found that six gene families are highly duplicated only in *P. bursaria* (Fig. 2b). The WD40 domain occurs in a large group of gene families

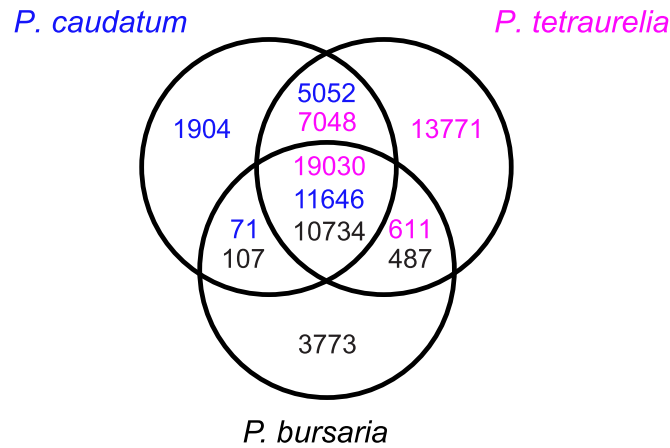
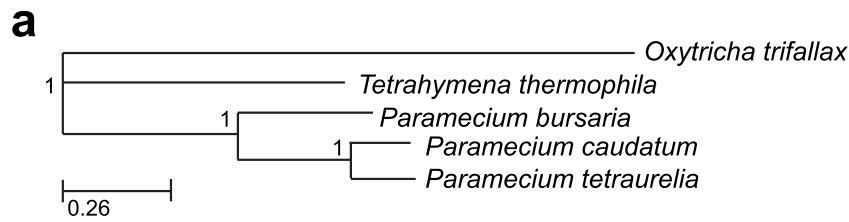
in eukaryotes, and WD40 domain-containing proteins often function as a scaffold in various cellular pathways. Moreover, WD40 domain-containing proteins have been reported to be involved in legume-*Rhizobium* symbioses [69]. TLDC domain-containing proteins primarily function as antioxidants to protect cells from reactive oxygen species (ROS) [70, 71], and algal endosymbiosis has been shown to generate excess ROS in different hosts [72]. Expansion of the TLDC domain-containing gene in *P. bursaria* may be an adaptation to such oxidative stress. Expression of *Paramecium* surface antigen domain-containing genes is often induced by environmental changes [73, 74], but it remains unclear why *Paramecium* surface antigen-containing, UPF0047 domain-containing, and protein kinase domain-containing genes are highly duplicated in *P. bursaria*.

### High heterogeneity of macronuclear minichromosome structures

Global DNA rearrangements during the development of new MACs are a peculiar feature of ciliates. Both precise elimination of IESs and imprecise removal of repeat regions have been reported for the *Paramecium* genus [34, 75]. After imprecise elimination of DNA, chromosome ends are either joined with other DNA fragments or telomeric repeats are added de novo. Taking advantage of long-read sequencing, we identified about 9000 reads with telomeric repeats ( $(C_4A_2)_n$  or  $(T_2G_4)_n$ ) at both ends of a single read, which represent full-length minichromosomes. The size distribution of minichromosomes is concordant with the chromosome sizes observed in the pulsed-field gel electrophoresis (Fig. 1a and Fig. 3a; Additional file 1: Fig. S3 [76, 77]). Moreover, the telomere addition sites were highly variable even for minichromosomes carrying similar gene contents (Fig. 3b; Additional file 1: Fig. S3d), which could further contribute to non-uniform gene dosage. Alternative telomere addition sites have been observed in other *Paramecium* species [78–80]. However, the *P. bursaria* macronuclear genome exhibited a quantitative difference in that chromosome breakage sites are much denser, resulting in shorter minichromosomes and higher variation in copy number.

To rule out the possibility that this pattern is an artifact of the PacBio sequencing platform, we mapped the Illumina reads that ended with telomeric repeats ( $n = 99,428$ ) against the MAC genome assembly and found that the read depth was indeed correlated with those inferred from PacBio long reads ( $n = 144,615$ ; Spearman correlation  $\rho = 0.57$ ,  $p < 0.0001$ , Methods, Additional file 1: Fig. S4 and Additional file 4: Table S6). Moreover, the non-uniform pattern of gene dosage was further supported by the read depth of total Illumina reads (Spearman correlation  $\rho = 0.75$ ,  $p < 0.0001$ , Fig. 3c). These results indicate that alternative telomere addition





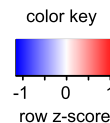
**b**

*P. bursaria*-specific protein family expansion

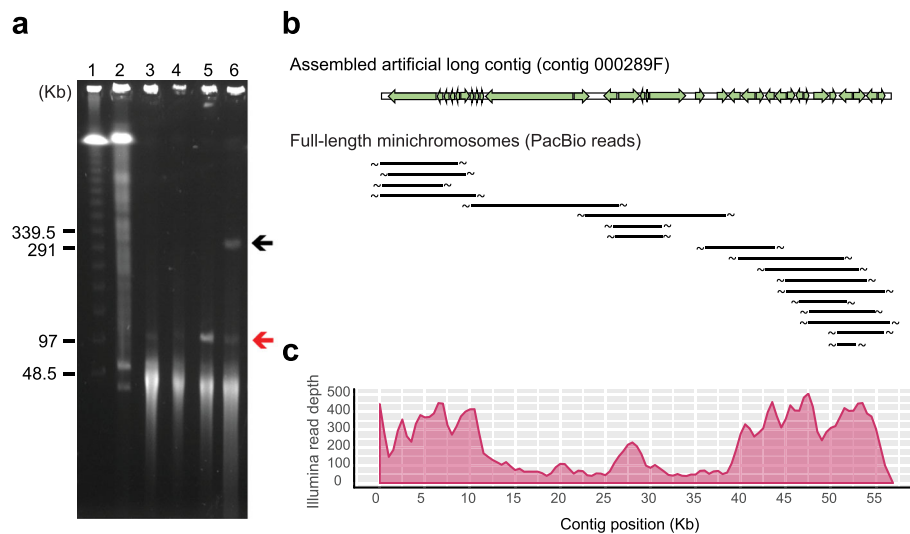
2	1	21	IPR017106 Coatomer_gsu, Coatomer gamma subunit
<i>P. t.</i>	<i>P. c.</i>	<i>P. b.</i>	

*P. bursaria*-specific gene duplication

0	5	30	IPR006571 TLDc_dom, TLDc domain
2	3	46	IPR002895 Paramecium_SA, Paramecium surface antigen
0	0	17	IPR001602 UPF0047, Uncharacterised protein family UPF0047
0	0	21	IPR017106 Coatomer_gsu, Coatomer gamma subunit
1	1	111	IPR036322 WD40_repeat_dom_sf, WD40-repeat-containing domain
2	0	29	IPR000719 Prot_kinase_dom, Protein kinase domain
<i>P. t.</i>	<i>P. c.</i>	<i>P. b.</i>	



**Fig. 2** Comparison of different *Paramecium* genomes. **a** Seventy-five percent of *P. bursaria* genes have orthologs in other *Paramecium* species. The phylogenetic tree was constructed using the core gene set of ciliates listed in Table 1 (see the “Methods” section). Numbers beside each node are posterior probability values obtained from Bayesian posterior probability. The scale bar represents the number of amino acid substitutions per site. The Venn diagram shows the number of orthologous genes shared among three *Paramecium* species. Orthologous genes were defined by comparing all protein sequences using OrthoFinder [61]. Different colors indicate the gene number of each species. The gene number of *P. tetraurelia* is more than that of other species due to the species having undergone two whole-genome duplication events [47]. **b** Heat map of the protein families and duplicated genes enriched in the *P. bursaria* genome. The gene numbers in each protein family/group of paralogs of each species are shown. The proportion of genes belonging to each protein family/group of paralogs to the whole genome gene number in each species was converted into a z-score by z-transformation ( $z = (x - \mu) / \sigma$ , where  $\mu$  is the population mean and  $\sigma$  is the population standard deviation). A two-proportional test was conducted and proteins or gene families are shown having a  $p$  value  $\leq 0.05$  after Benjamini-Hochberg adjustment



**Fig. 3** High heterogeneity of minichromosomes leads to non-uniform gene dosage. **a** The length distributions of minichromosomes are similar between young and old cells. Pulsed-field gel electrophoresis (PFGE) patterns of the aposymbiotic cells of Dd1 (lane 3), KM2 (lane 4), DK1 (lane 5), and the symbiotic cells of newly formed progeny DK2 (lane 6). Both DK1 and DK2 are the progeny of Dd1 and KM2 and they are about 250 and 17 generations old, respectively. The red arrow indicates the mitochondrial DNA and the black arrow indicates the *Paramecium bursaria* *Chlorella* virus-1 (PBCV-1) DNA of the green algae (see also Fig. S3 [see Additional file 1]). Lane 1 is the Lambda DNA ladder. Lane 2 is the control sample of the *T. thermophila* Bll minichromosomes to show that full-length minichromosomes were maintained during our sample preparation. **b** An example showing that minichromosome copy numbers are highly variable. In this case, long reads with telomere repeats at both ends from the PacBio data have been mapped to an assembled artificial long contig 000289F. Each read represents an intact minichromosome in the MAC. Green arrows represent the location and orientation of predicted coding sequences. The ~ symbols represent telomeres of minichromosomes. **c** The read depth of the Illumina data is highly correlated with PacBio data (Spearman correlation  $\rho = 0.75$ ,  $p$  value  $< 0.0001$ ). Sequencing coverage is calculated from the depth of Illumina reads. The distribution of read depth has been drawn using a 1-kb interval and 0.5-kb sliding window. Genomic DNA of the Dd1 strain was used to generate the PacBio and Illumina data

occurs throughout the macronuclear genome of *P. bursaria* (Additional file 1: Fig. S5).

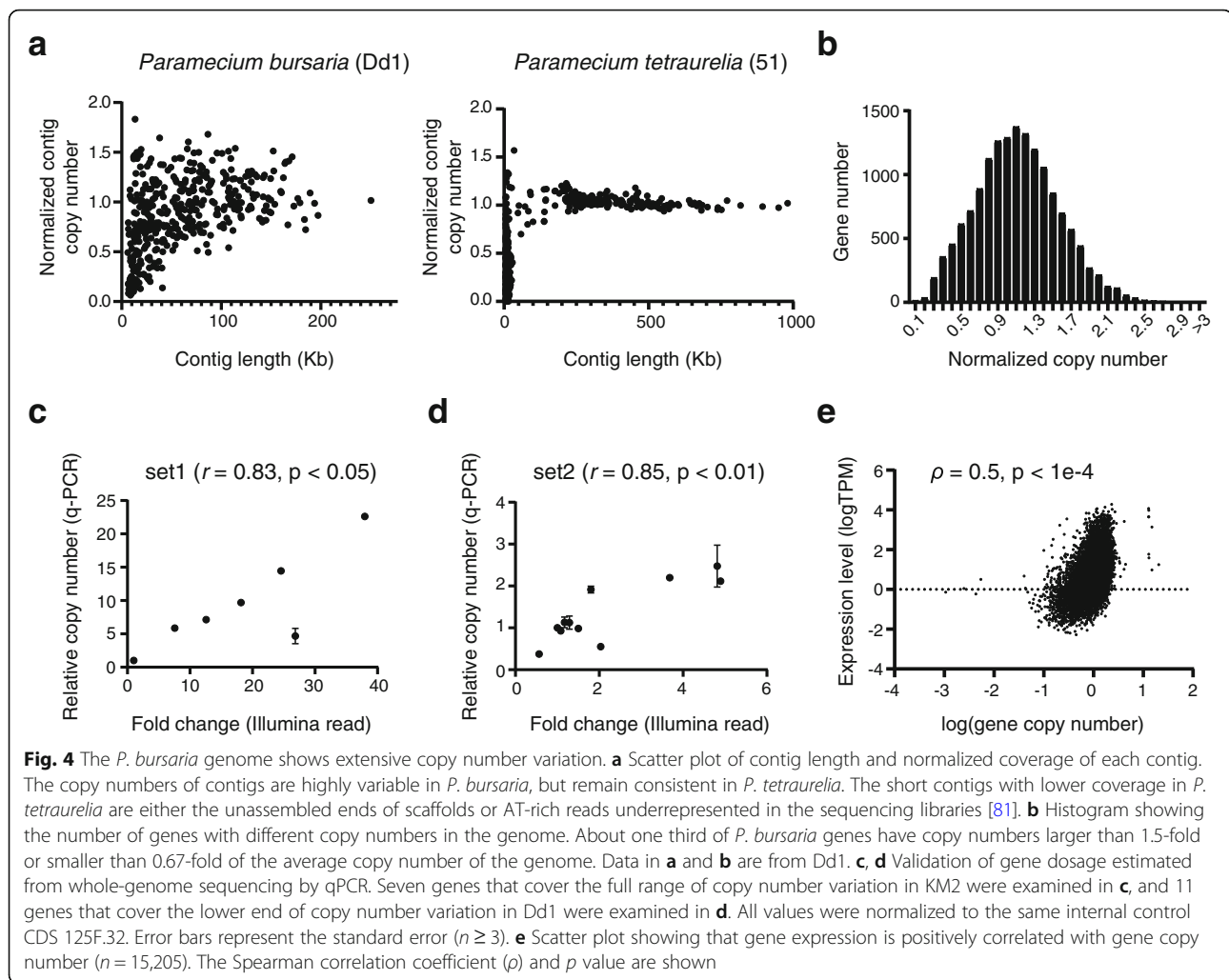
Previously, clonal aging has been shown to cause random fragmentations of minichromosomes in *P. tetraurelia* [35]. Although clonal aging has not been reported in *P. bursaria*, there was a possibility that the short and highly variable minichromosomes observed in our strains were the outcome of long-term asexual propagation. To investigate this possibility, we examined the MAC genomes of two young F1 diploids, DK1 and DK2, generated from crossing the Dd1 and KM2 strains. The DK1 cells had been asexually propagated for about 250 generations and their endosymbionts had been removed. DK2 cells were immature young cells (about 17 generations old) that still harbor green algae. The pulsed-field electrophoresis data showed that both young and old cells exhibited similar length distributions of minichromosomes (Fig. 3a; Additional file 1: Fig. S3d). Moreover, whole-genome sequencing data revealed that young DK2 cells also exhibit highly variable breakage sites (Additional file 1: Fig. S6). These data rule out the possibility that the high heterogeneity of minichromosome structures observed in our *P. bursaria* strains is caused by clonal aging. Moreover, it indicates that variable chromosome

breakage occurs during or soon after the development of new MACs in sexual reproduction.

#### Extensive copy number variation in the macronuclear genome

The copy number of individual minichromosomes in *T. thermophila* and *P. tetraurelia* is uniform [31, 81]. When we analyzed the average read depth of assembled contigs, a majority of them fell within a range of 0.67-fold to 1.5-fold of the genome average (Fig. 4a; Additional file 1: Fig. S6a). In contrast, the average read depth of *P. bursaria* contigs was distributed across a broader range (Fig. 4a). CNV could be further depicted at the gene level. In the Dd1 reference genome, about one third of *P. bursaria* genes (5249/15,101 = 35%) had copy numbers larger than 1.5-fold or smaller than 0.67-fold of the average copy number of the genome (Fig. 4b). Similar trends were also observed in other *P. bursaria* strains, including the newly generated DK2 cells (Additional file 1: Fig. S6b).

To confirm that CNV was not due to a bias in mapping or sequencing processes, we selected two sets of genes to validate copy numbers by quantitative PCR. The first set comprised genes covering the full range of CNV and the second set represented only genes with



lower copy numbers. Our results showed high correlations between quantitative PCR and Illumina sequencing data (Fig. 4c and d), indicating that *P. bursaria* can tolerate a wide range of gene dosages, unlike *T. thermophila* and other sequenced *Paramecium* species.

In other organisms, changes in gene copy number are often associated with alterations in gene expression levels [3, 18, 82]. We observed a positive correlation between mRNA abundance and gene copy number in *P. bursaria* (Fig. 4e), suggesting that in addition to transcriptional regulation, the observed CNV also contributes to variation in gene expression.

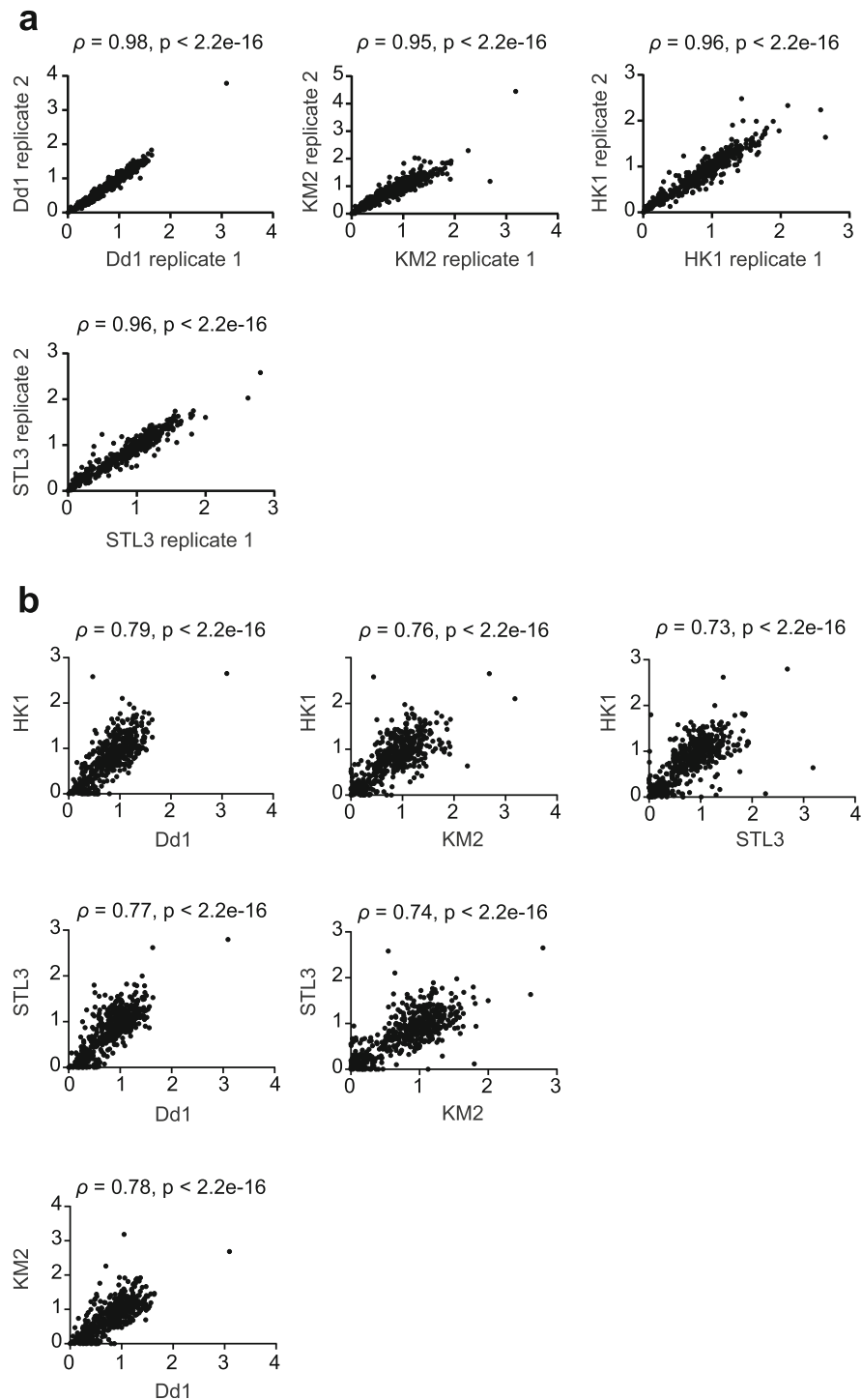
#### Gene copy number variation is correlated with the gene function

Random minichromosome assortment has been suggested to be the primary mechanism for chromosome segregation in MAC during asexual reproduction of ciliates. Amitosis is poorly characterized and it remains unclear if the patterns of CNV can be stably maintained during amitosis. To address this question, we sequenced

and compared genomic DNA from different replicate populations of the same strains. Contig copy numbers were highly correlated between replicates even though some of the populations had diverged for more than 100 generations (Fig. 5a; Additional file 2: Table S9). We observed similar patterns when we compared the copy number of individual genes (Additional file 2: Table S7). The slightly reduced correlation is probably due to the small size of genes that makes the data noisier. These data suggest that CNV patterns can be stably maintained during asexual reproduction.

More interestingly, when we compared the copy numbers of assembled contigs between genetically divergent strains, the correlation coefficients remained high but they were significantly lower than those obtained from within-strain comparisons (Fig. 5b) ( $p < 0.005$ , one-tailed Mann-Whitney  $U$  test). One possible explanation for this outcome is that genes related to basic cellular pathways have conserved CNV patterns between different strains, but genes related to strain-specific or condition-specific phenotypes have more variable CNV patterns.





**Fig. 5** The copy numbers of contigs are conserved between replicates and different strains. **a** Contig copy numbers are strongly correlated between two biological repeats of the same strain collected at different time points. **b** The correlation of contig copy numbers between different strains is still high, but lower than that between two replicates. The  $x$ - and  $y$ -axes represent the read depth of each contig normalized to both contig length and whole-genome coverage for the respective strain. The Spearman correlation coefficient ( $\rho$ ) and  $p$  value are shown

To test this hypothesis, we calculated between-strain copy number variability for each gene using the genomic data from four different *P. bursaria* strains (Fig. 6a). The

top (non-conserved group) and bottom (conserved group) 25% of genes in terms of their CNV were chosen for subsequent analyses.

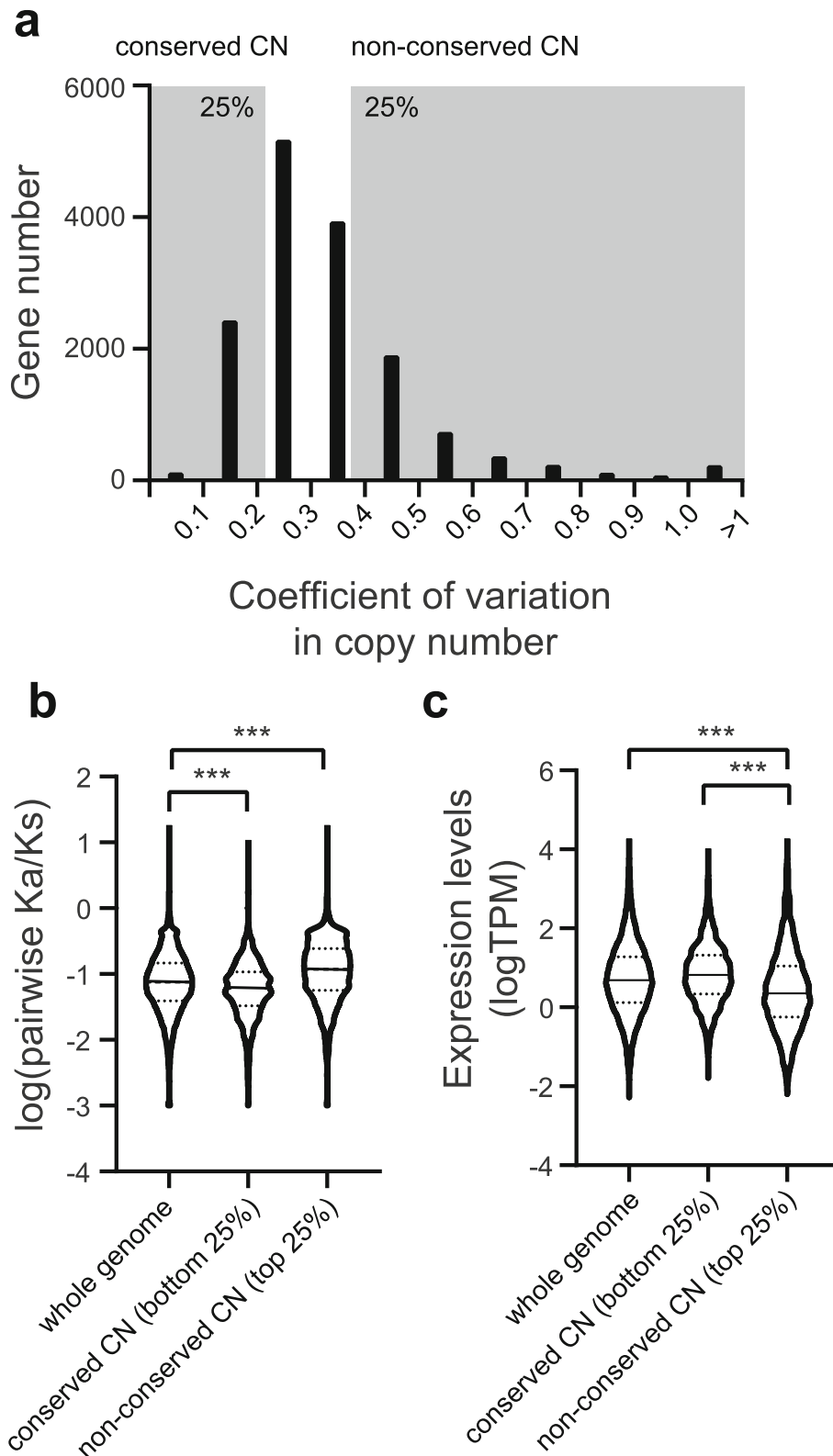


Fig. 6 (See legend on next page.)

(See figure on previous page.)

**Fig. 6** Copy number variation is correlated with evolutionary rate and expression levels. **a** The distribution of coefficient of variation (CV) in copy number for each gene using population genomics data from four strains. The top and bottom 25% of genes in terms of CV (labeled in gray) are categorized into non-conserved and conserved copy number (CN) groups, respectively. **b** Pairwise Ka/Ks values are significantly lower in the conserved CN group and higher in the non-conserved CN group when compared to those of the whole genome. **c** Gene expression of the non-conserved CN group is significantly lower than that of the conserved CN group. \*\*\* $p$  value < 0.0001, two-tailed Mann-Whitney  $U$  test

First, we examined the relative evolutionary rate (Ka/Ks) of these two groups of genes. In general, genes involved in basic cellular functions or occupying a central position in regulatory networks have lower evolutionary rates, whereas genes involved in strain-specific or condition-specific phenotypes evolve rapidly [83–85]. We found that the evolutionary rate of the conserved group of genes was significantly lower than that of the whole genome gene set ( $p < 0.0001$ , two-tailed Mann-Whitney  $U$  test, Fig. 6b). In contrast, the non-conserved group had a significantly higher Ka/Ks value than the whole genome gene set ( $p < 0.0001$ , two-tailed Mann-Whitney  $U$  test). Second, we performed GO enrichment analysis. The conserved group tended to be involved in general pathways, such as basic cellular processes or cellular component biogenesis (Additional file 2: Table S8), whereas the non-conserved group was enriched for functions related to environmental responses. Finally, we compared the gene expression levels of these two groups of genes. Housekeeping genes are stably expressed, whereas genes related to condition or strain adaptation are often only induced under specific conditions [86]. Indeed, we found that the non-conserved group presented significantly lower expression than either the conserved group or the whole genome dataset ( $p < 0.0001$ , two-tailed Mann-Whitney  $U$  test, Fig. 6c). Together, our data suggest that *P. bursaria* has evolved the ability to adjust gene copy number according to gene function.

## Discussion

Due to high heterozygosity and structural variation of the two haplotypes in the diploid cell, it is notoriously difficult to annotate correctly the macronuclear genome of *P. bursaria*. So far, no homozygous *P. bursaria* diploid lines have been established since *P. bursaria* does not undergo autogamy (an autofertilization process) like other *Paramecium* species [87]. In a recent study reporting a draft of the *P. bursaria* genome [88], only about 7000 genes had orthologs in closely related *P. caudatum*, despite 17,226 genes being annotated. By combining population genomic data and gene expression profiles, we have established a functional genome of *P. bursaria* that contains 15,101 genes, 72% (10,841/15,101) of which share orthologs with *P. caudatum* (Fig. 2a). Our significantly improved annotation allowed us to perform more in-depth genome comparisons between *Paramecium* species and for intraspecific populations.

*P. bursaria* has very short intergenic and intronic regions compared to other characterized eukaryotic genomes [31, 59, 89, 90]. Our nucleotide diversity analysis indicated that the intergenic sequences are most conserved in the first 50 bp upstream of the translation start site (Fig. 1c), suggesting that transcription regulatory elements are mainly located within this small region. Compared to other eukaryotic organisms that possess sophisticated regulatory elements in long intergenic regions, the functional flexibility of *cis*-regulatory elements in *P. bursaria* may be restricted. This raises the possibility that the CNV adjustment provides another layer of flexibility for gene regulation. Despite the 5' and 3' splice sites being highly conserved in our annotated genes, the *P. bursaria* genome only contains orthologs of 72% and 79% of splicing-related genes in yeast and plants, respectively. Moreover, several essential splicing-related genes shared by both yeast and human cells are missing from the *P. bursaria* genome. We extended our analysis to other ciliate genomes, including other *Paramecium* species and *T. thermophila*, and confirmed that these splicing-related genes are absent from these ciliates. This outcome indicates that these essential proteins have been substituted by other uncharacterized functional orthologs or the splicing machinery has been extensively modified in ciliates. Since every *P. bursaria* gene contains 2.4 introns on average (Additional file 2: Table S3), the splicing machinery is not a “reduced” form of spliceosomes as observed in intron-poor eukaryotes [91] even if it significantly diverges from that which operates in other eukaryotic cells. Ciliates provide an interesting system for studying the evolution of spliceosomes.

We identified 102 internal structural variations (ISVs) that account for 2% of our genome assembly, and 238 genes were annotated from ISVs. These ISVs were generated from assembled contigs that shared a significant portion of homologous regions with other contigs (Additional file 1: Fig. S1b). To avoid preserving redundant contigs (and CDSs) in the assembly, we only retained the unique regions that contained CDSs. There are two possible origins for those ISVs. First, an ISV may represent one of a pair of homologous chromosomes carrying some highly diverged sequences. In this scenario, the ISV would only have one haplotype in the genome. We counted 71 ISVs that belong to this type. In contrast, 31 ISVs were observed to have two haplotypes, suggesting that the redundant contigs may be isoforms generated through

different DNA rearrangements of the same micronuclear chromosome. Alternative DNA rearrangements have been observed in other ciliates. For instance, an alternative DNA deletion of the M element in *T. thermophila* results in two minichromosomes with different lengths in the MAC [92]. Studies in *O. trifallax* have further shown that extensive alternative DNA rearrangements can generate minichromosomes with different gene products [93–95]. Whether alternative DNA rearrangements have specific regulatory functions in *P. bursaria* remains to be addressed.

Our data reveal intricate chromosome breakage patterns in the MAC, which further contribute to highly variable gene dosages in the *P. bursaria* genome. Although variable gene dosages were also observed in *Spirotrich*, the composition and generation of macronuclear chromosomes are different from *P. bursaria*. For instance, the majority of the nanochromosomes of *O. trifallax* contain only one gene and the telomere addition sites are primarily located in the intergenic regions [36]. We found most of the minichromosomes in *P. bursaria* contain more than one gene and ~30% of the telomere addition sites are located in the coding regions. Interestingly, the cluster analysis indicated that the proportion of genes in clusters was significantly higher in the conserved copy number group (79.0%) compared to random sampling (45.1%, 95% confidence interval 43.4–46.8%,  $p$  value < 0.001, bootstrapping = 1000, see the “Methods” section). It raises the possibility that conserved genes may be clustered on the same minichromosomes so the copy number can be adjusted together.

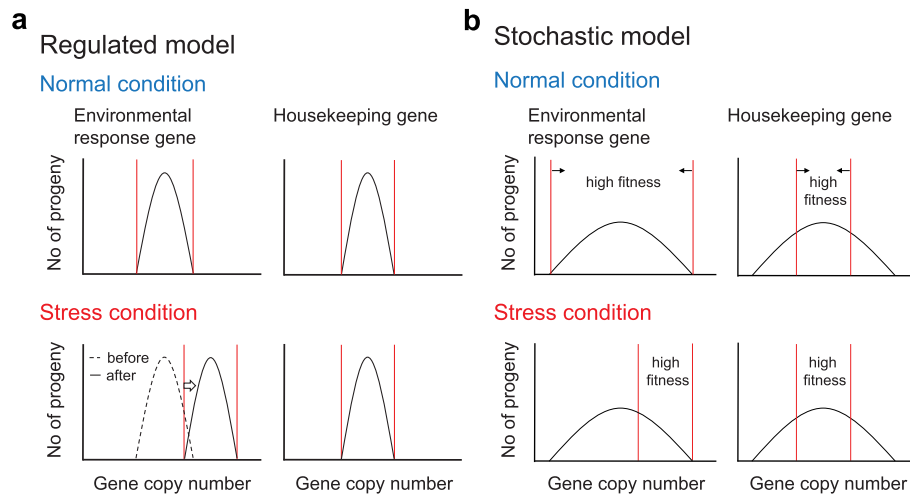
In other ciliates, specific motifs have been identified near the chromosome breakage sites only when precise elimination is involved [27–29, 96, 97]. A study of *P. tetraurelia* suggested that chromosome breakage sites are often located near repeated elements or transposons, but no specific motif was found [34]. We compared the distributions of telomere addition sites between different *P. bursaria* strains and observed significant correlations (Spearman correlation  $\rho = 0.51$ ,  $p < 0.0001$  for KM2 and Dd1;  $\rho = 0.54$ ,  $p < 0.0001$  for DK1 and Dd1;  $\rho = 0.52$ ,  $p < 0.0001$  for KM2 and DK1). It suggests that different strains share some chromosome breakage hotspots and sequence motifs may be involved. Since the breakage sites in *P. bursaria* exhibit high heterogeneity, we used genomic regions that have at least four telomere-containing reads to search for specific motifs (see the “Methods” section). Only GA-rich motifs were found to be highly enriched (Additional file 1: Fig. S7 [98]). More experiments are required to validate the cis-regulatory functions of these motifs. Nonetheless, our findings should prompt further investigations of the molecular mechanisms underlying complex chromosome breakage.

The correlation between gene copy numbers and expression levels reveals the biological impact of CNV, as previously observed in other organisms [11, 13–15, 99]. The distinct evolutionary rates and expression levels between highly variable and lowly variable gene groups further suggest that CNV levels may be adjusted or selected according to gene function. At least two possible scenarios can explain our current observation. First, the copy numbers of individual minichromosomes are always controlled in a specific range according to their functions. When the cells encounter environmental changes, the copy numbers of minichromosomes required for adapting to the new environment are also adjusted (Fig. 7). In *Oxytricha*, an RNA-mediated mechanism has been shown to regulate chromosome copy number during sexual reproduction [21]. A similar but modified mechanism may be used by the *P. bursaria* cells to adjust CNV during asexual reproduction. As an alternative scenario, replicated minichromosomes are randomly distributed to daughter cells during amitosis. However, only the progeny carrying correct dosages of housekeeping genes will survive, whereas copy numbers of non-housekeeping genes are allowed to drift unless they are required for a specific condition (Fig. 7). Under this scenario, the CNV pattern is shaped by natural selection and the cells do not develop specific mechanisms to adjust it. Although the second scenario sounds much more straightforward compared to the first one, it imposes a high fitness cost on the whole population. Through random segregation of minichromosomes, a certain proportion of daughter cells are likely deemed unfit since they do not have the correct composition of dosage-sensitive genes. More experiments are required to resolve this issue.

Why does *P. bursaria* exhibit and tolerate such a broad range of CNV? In other eukaryotic organisms, such large-scale CNV regulation can only be observed in asexual populations when cells encounter drastic environmental challenges [100, 101]. It is known that the genetic material of MACs is specific to the asexual life cycle of ciliates and that new MACs are generated during each sexual cycle. Is it possible that the high genome plasticity observed in *P. bursaria* represents a general and effective strategy to regulate gene dosage and expression without complex transcriptional regulatory networks? Further investigation of the ciliate genome will help us understand the general rules of how high genome plasticity is achieved and tolerated, especially during acute stress conditions.

## Conclusions

Unlike other sequenced *Paramecium* species, the macronuclear genome of *P. bursaria* exhibits a wide range of gene copy number variation. We analyze the patterns of CNV between different populations of *P. bursaria* and reveal that the CNV patterns are partially conserved.



**Fig. 7** Models explaining how the copy number of genes with different functions is adjusted or selected. **a** In the regulated model, the duplicated minichromosomes distributed to the daughter cells are controlled within a range during amitosis. When cells encounter a stress environment, the copy number of environmental response genes is specifically adjusted to adapt to the new environment. **b** In the stochastic model, the duplicated minichromosomes are randomly distributed to the daughter cells during amitosis and the CNV pattern is shaped by natural selection. The housekeeping genes are more sensitive to dosage variation so only the progeny carrying correct dosages display high fitness and there is no selection for the dosage of environmental response genes under normal conditions. However, cells carrying certain copies of environmental response genes would be selected for in a stress environment

The group of genes with consistent CNV patterns among populations comprises sequence-conserved genes with housekeeping functions, whereas the group with variable CNV patterns often includes environmental response or species-specific genes. We further show that mRNA levels are partially correlated with gene copy number. Our data suggest that *P. bursaria* exhibits high plasticity and tolerance to gene copy number variation, which may play a general role in allowing cells to adapt to different environments.

## Methods

### Strains and culture conditions

The *P. bursaria* strains Dd1, KM2, HK1, and STL3 were obtained from the Symbiosis Laboratory, Yamaguchi University (<http://nbrpcms.nig.ac.jp/paramecium/?lang=en>), with partial support from the National Bio-Resource Project of the Japan Agency for Medical Research and Development. The endogenous endosymbiont in all of those strains is *Chlorella variabilis*. Original collection locations and dates of all strains are presented in Table S9 [see Additional file 2]. The DK1 and DK2 strains are descendants of Dd1 and KM2. To generate the progeny of Dd1 and KM2, aposymbiotic cells of Dd1 and symbiotic cells of KM2 in the early stationary phase were collected and washed twice with modified Dryl's solution (in which  $\text{KH}_2\text{PO}_4$  was used instead of  $\text{NaH}_2\text{PO}_4 \cdot 2\text{H}_2\text{O}$ ). The cells were starved for 2 days and several hundred cells were mixed. After 5 h, each of the conjugating pairs was transferred to individual

modified Dryl's solution droplets. After the two exconjugants separated, we immediately isolated each of them to establish an independent clone. The clones were fed with *K. pneumoniae*-infusion lettuce medium (2.5% Boston lettuce juice in modified Dryl's solution) after at least 2 days to avoid macronuclear regeneration. After the F1 cells started propagating, we used PCR followed by Sanger sequencing to check the genotype of specific regions in the progeny. If the genotype was a combination of both parents, then the cells were used for further analyses. The primers used for PCR are listed in Table S10 [see Additional file 2]. Both DK1 and DK2 cells showed a period of sexual immaturity after conjugation, which is consistent with the phenotype of true cross-fertilized clones. Lastly, we analyzed the whole-genome sequencing data and compared unique sequence variations (see the "Whole-genome analysis" section). Both DK1 and DK2 contained a proportion of, but not all, unique sequence variations from Dd1 and KM2 (Additional file 1: Fig. S8), providing direct evidence that DK1 and DK2 were real F1 progeny of Dd1 and KM2. The sexually reproduced strains DK1 and DK2 were generated on 25 November 2016 and 15 August 2019, respectively.

*P. bursaria* cells were grown in lettuce medium and fed with *Klebsiella pneumoniae* (NBRC 100048 strain). The cell cultures were kept at 23 °C with a light to dark cycle of 12 h:12 h. Fresh bacteria-containing lettuce medium was added every 2 days. The cells entered early stationary phase 1 day after their last feed.



To obtain aposymbiotic *Paramecium* strains, green symbiotic cells were treated with cycloheximide (10 µg/ml) as previously described [102].

#### Genomic DNA preparation

Aposymbiotic cells of Dd1 in early stationary phase were starved for two further days and collected using a filtering apparatus with 11-µm-pore-size nylon membrane (NY1102500, Millipore, Burlington, MA, USA). To isolate macronuclei,  $\sim 2 \times 10^6$  cells were washed twice with modified Dryl's solution and lysed using an equal volume of 0.25 M TCMS buffer (10 mM Tris-HCl pH 8.0, 2 mM CaCl<sub>2</sub>, 8 mM MgCl<sub>2</sub>, 0.25 M sucrose) with 0.3% (w/v) NP-40. Cell extract was sonicated ( $\sim 100$  watts; Sonicator 3000, Misonix Inc., Farmingdale, NY, USA) for 6 min to dissociate the macronuclei and micronuclei. We carefully loaded 7 ml of the extract onto 8 ml of 1.6 M TCMS buffer in a 15-ml falcon tube and centrifuged at 1800 rcf and 4 °C for 15 min. The macronuclear pellet was washed once with 10 ml 0.25 M sucrose and again centrifuged at 1800 rcf and 4 °C for 15 min. The precipitate was collected and a small amount was stained with SYTOX Green to calculate the ratio of macronuclei and micronuclei. Only precipitates with a ratio of MAC number to total nucleus number > 0.8 were used. A previous study reported the ratio of DNA content of micronucleus and macronucleus in *P. bursaria* to be approximately 1:23 [52]. After we had enriched for MAC, reads derived from MIC were low enough to be discounted. The DNA was extracted using QIAGEN Genomic-tip 20/G (Cat No.10223, QIAGEN, Venlo, Netherlands).

To isolate total genomic DNA from whole cells,  $\sim 1 \times 10^6$  cells were washed twice with modified Dryl's solution and lysed by an equal volume of 0.25 M TCMS buffer with 0.3% (w/v) NP-40. Genomic DNA was extracted using QIAGEN Genomic-tip 20/G (Cat No.10223, QIAGEN).

#### RNA preparation and data analysis

Approximately  $1 \times 10^5$  Dd1 cells in early stationary phase were collected using a filtering apparatus with 11-µm-pore-size nylon membrane and washed twice with modified Dryl's solution. Total RNA was extracted using TRI Reagent (T9424, Sigma-Aldrich) and RNeasy Mini kit (Cat No. 74106, QIAGEN) following the manufacturer's instructions.

RNA-seq libraries were prepared with the Illumina TruSeq Stranded mRNA LT Sample Prep Kit (Illumina, San Diego, CA, USA), and sequenced using an Illumina Next-seq platform. The reads were quality-trimmed using Trimmomatic v0.36 with options ILLUMINA\_CLIP = 2:30:10, LEADING = 3, TRAILING = 3, SLIDINGWINDOW = 4:15, and MINLEN = 36. The transcripts per million (TPM) for each sample was quantified using Salmon v0.9.1 with option numBootstraps = 200; gcBias [103].

#### Pulsed-field gel electrophoresis and southern blot

Cells of different strains and species ( $\sim 1.5 \times 10^4$  cells per plug for *P. bursaria*,  $\sim 5 \times 10^5$  cells per plug for *Tetrahymena thermophila* BII) were washed twice in ET buffer (100 mM EDTA, 10 mM Tris buffer [pH 8.0]) and suspended in 100 µl ET buffer. The liquid was mixed with 200 µl 1.5% low gelling temperature agarose (A9414, Sigma-Aldrich, St. Louis, MO, USA) and solidified in a casting mold (Biometra, Göttingen, Germany). The agarose plugs were incubated in lysis solution (10 mM Tris buffer [pH 8.0], 1% SDS, and 1 mg/ml proteinase K, 0.1 M EDTA) overnight at 50 °C and washed for 1 h three times in TE buffer (50 mM EDTA, 10 mM Tris, pH 8.0) at room temperature. Subsequently, the agarose plugs were inserted into the well of a 1% gel (A2929, Sigma-Aldrich) to perform pulsed-field gel electrophoresis in the CHEF DR II system (Bio-Rad, Hercules, CA, USA) (0.5X TBE, ramping switch time of 2.1 s to 54.1 s, 6 V/cm, angle 120° for 17 h). The Lambda Ladder-CHEF DNA Size Standard (#170-3635, Bio-Rad) was used as a size marker. Visualization was performed after staining with ethidium bromide.

Total genomic DNA of the aposymbiotic DK1 strain was labeled with Digoxigenin (DIG) by using DIG-High Prime (Cat. No. 11585606910, Sigma-Aldrich) and the specific probes for mtDNA, the PBCV-1 genome, and gene 000334F\_H1.12 (a gene encoding the homolog of alpha-tubulin) were generated using PCR DIG Probe Synthesis Kit (Cat. No. 11636090910, Sigma-Aldrich). The DNA fragments separated on the agarose gel were transferred to an Immobilon-NY+ nylon membrane (INCY00010, Millipore) and hybridized with different probes at 42 °C in the DIG Easy Hyb buffer prepared from DIG Easy Hyb granules (Cat. No. 11796895001, Sigma-Aldrich). After overnight hybridization, the membrane was washed twice with 2X SSC and 0.1% SDS at 68 °C for 15 min and then washed twice with 0.5X SSC and 0.1% SDS at 68 °C for 15 min. For immunological detection, the membrane was incubated in 2X blocking buffer (2% Casein, 0.1 M Maleic acid, 0.15 M NaCl, pH 7.5) at 37 °C for 30 min and incubated with anti-DIG-AP, Fab fragment (1:10000, 11093274910, Roche Applied Science, Penzberg, Germany) at 37 °C for 30 min. The membrane was washed twice with washing buffer (0.1 M maleic acid, 0.15 M NaCl, 0.3% Tween 20, pH 7.5) at 37 °C for 15 min and equilibrated in detection buffer (0.1 M Tris, 0.1 M NaCl, 0.05 M MgCl<sub>2</sub>, pH 9.5) at 37 °C for 5 min. The hybridization fragments were detected using CSPD-ready to use (Cat. No. 11755633001, Sigma-Aldrich) according to the manufacturer's instructions.

#### Genome sequencing and assembly

A 20-kb insertion genomic DNA library was prepared according to a standard protocol and sequenced using the PacBio Sequel platform (Pacific Biosciences, Menlo

Park, CA, USA) of the NGS core of Academia Sinica (<http://ngs.biodiv.tw/NGSCore/>) to obtain long reads data. Paired-end libraries were prepared by a standard protocol and sequenced using the Illumina Miseq and Nextseq platforms of the Genomics core of IMB (<http://www.imb.sinica.edu.tw/mdarray/>). The reads were quality-trimmed using Trimmomatic v0.36 with options ILLUMINACLIP = 2:30:10, LEADING = 3, TRAILING = 3, SLIDINGWINDOW = 4:15, and MINLEN = 36. The subreads were assembled using Falcon v.0.4.0 with options overlap minlen = 1000, to graph minlen = 1000, max\_diff = 100, max\_cov = 100 and min\_cov = 5 [104], and the sequence was polished by Illumina reads using Pilon v1.22 [105].

We assembled 614 contigs and removed 17 contigs derived from bacterial contamination judging from their GC content and similarity to bacterial sequences according to the NCBI database. Many diploid genomes contain regions with high levels of heterozygosity, which increases the difficulty of distinguishing two haplotypes [106, 107]. If the regions with highly divergent alleles are assigned as two alternate contigs, the assembled genome will become fragmented and bigger than the real genome. We used Redundans v0.11 [108] to generate the haploid genome. The default criteria for Redundans (established from simulation and real data by the developers) to determine alternative contigs as haplotypes are as follows. First, the alignment length should be more than 80% of the shorter contig. Second, the sequence identity should be greater than 50%. To increase stringency in our analysis, we used 70% as a cutoff for the second criterion. We used MUMMER v3.23 [109] to generate the haploid genome and manually joined the overlapping contigs. To distinguish the two haplotypes, we identified variations by GATK v4.0.3 [110] with default settings and filtered out the variants located in the regions with lower depths ( $\leq 30$ ) or had a lower variant read depth ( $\leq 10$ ). Among the 661,718 heterozygous sites identified between H1 and H2 haplotypes of Dd1, only 40 sites have more than two genotypes. The results suggested that the majority of our assembly did not collapse duplication regions into a single locus. We constructed the diploid genome by HAPCUT2 v1.0 [111] according to the variant information acquired from the GATK algorithm. The effect of variants was examined by SNPEFF 4.4 [112]. The accession numbers of each sequencing dataset can be found in Table S9 (see Additional file 2). The mapping rates for the reads from RNA-seq and DNA-seq were 96.4% and 95.6%, respectively, suggesting that our assembly represents most of the functional genome.

### Genome annotation

Gene annotation was performed by comparing our *P. bursaria* RNA-seq data generated from the Dd1 strain

and protein sequences from *P. caudatum* and *P. tetraurelia* acquired from ParameciumDB (<http://paramecium.i2bc.paris-saclay.fr>) using the pipeline described in Arnaiz et al. [113]. We annotated the genes for both haplotypes and then used the following criteria to select for the functional gene set of *P. bursaria*. First, we aligned gene sequences by BWA-MEM [114, 115] and grouped alleles from the same locus in the four different sequenced *P. bursaria* strains. If the alleles shared the same sequence positions for the start and stop sites when aligned, they were classified into a CDS subgroup (Fig. 1b). For each locus, we chose the subgroup with the highest number of alleles to identify the functional allele. If the difference between the allele numbers of the top two subgroups was equal to or less than one, we chose the subgroup with a longer coding region. Once the subgroup was selected, we calculated alignment bit scores for all allele pairs using BLASTp and selected the allele with the highest average score as the representative functional gene. Potential protein domains of each annotated gene were identified using InterProScan 5.30–69.0 [116], and the gene was assigned to gene ontology terms according to its protein domains.

To assess the completeness of our genome assembly, we used the following criteria to filter the alignment of Core Eukaryotic Gene (CEG) sequences and our genome from BLASTP [56]. First, the alignment length had to cover over 70% of the aligned CEG sequence. Second, the alignment *E*-value should be less than  $1e-10$  [36].

### Calculation of nucleotide diversity $\pi$ of intergenic regions

We used CDS subgroups containing more than four alleles in the populations ( $n = 12,435$ ) to calculate the nucleotide diversity per site. We investigated the region upstream of the start codon of each gene until the point where it encountered the next gene, but with a maximum length of 500 bp. Nucleotide diversity was calculated using the formula  $\pi = 2pqn/(n - 1)$ , where  $p$  and  $q$  are the major and minor allele frequencies and  $n$  is the total number of alleles in each group [117].

### Construction of species phylogenetic tree

We identified homologs among different ciliate species using OrthoFinder v2.0.0 [61] with default settings. In each orthogroup, an individual species may contribute more than one ortholog if there is a recent gene or genome duplication event in that species. For example, *P. tetraurelia* often contributes multiple genes to each orthogroup since it has experienced two whole-genome duplication events. Some orthogroups may contain genes from only one species if they are species-specific genes. To draw the Venn diagram in Fig. 2, different types of orthogroups (e.g., the orthogroup containing orthologs from all species) were collected and the total gene

numbers from each species were counted. To construct the phylogenetic tree, we chose genes having only one ortholog in each species ( $n = 493$ ) to perform the analysis. The protein sequences of each orthogroup were aligned independently using MUSCLE v3.8.31 [118]. The conserved blocks from the multiple sequence alignment of each orthogroup were selected using Gblocks v0.91b [119] and concatenated. IQ-TREE v1.6.10 was used to select the best-fit model for sequence evolution (LG+I+G4) and to reconstruct the phylogenetic tree according to a Maximum Likelihood approach [120, 121]. The graph was drawn using the ETE toolkit [122].

### Gene ontology enrichment

GO categories with fewer than three genes were excluded. Enrichment scores were calculated by dividing the proportion of the genes of interest classified into the indicated category by the proportion of those genes in the genomic background. The significance of enrichment was analyzed using a hypergeometric test (phyper module) in R, and  $p$  values were adjusted using the Benjamini-Hochberg method [123].

### Identification of protein family and gene lineage expansions

The protein domains of each gene of *P. bursaria*, *P. caudatum*, and *P. tetraurelia* were identified using InterProScan 5.30–69.0 [116]. Each protein family contains both duplicated genes (which share high levels of sequence similarity) and non-duplicated genes that carry the same protein domain. To investigate the expanded protein families specific to *P. bursaria*, we calculated the gene number of each protein family in each species. A two-proportional test was used to test significance in three species, with a  $p$  value threshold equal to or smaller than 0.05 after Benjamini-Hochberg adjustment. To examine the orthologs specifically duplicated in *P. bursaria*, we categorized the genes in *P. bursaria*, *P. caudatum*, and *P. tetraurelia* into orthogroups depending on the protein similarity using OrthoFinder v2.0.0 [61] with default settings. A two-proportional test was used to test the orthogroups with a significantly different gene number in the three species, with a  $p$  value threshold equal to or less than 0.05 after Benjamini-Hochberg adjustment.

### Whole-genome analysis

For copy number variation (CNV) analysis, trimmed reads were mapped via BWA-MEM and read depth was analyzed using the SAMtools program [124]. The copy number of each contig was calculated by normalizing contig length and read depth to the whole genome coverage derived from the SAMtools bedcov module. The copy number of each gene was also calculated by

normalizing gene length and read depth to the whole genome coverage derived from the SAMtools bedcov module. The coefficient of variation (CV) for gene copy number was calculated using the copy numbers of each gene in all four strains.

We used a homemade Perl script to identify the Illumina reads that contained telomeric repeats. For Illumina data, the reads needed to contain at least 18 bp of telomere repeats with an allowance for one mismatch. For PacBio data, the reads needed to contain at least 30 bp of telomere repeats with an allowance of five mismatches. We used the same method described above for CNV analysis to calculate the read counts in 2-kb windows for both Illumina and PacBio reads (Additional file 4: Table S6). For the correlation analysis of coverage between two sets of data, the read depth in 2-kb windows was normalized to the window length to acquire the coverage, and Spearman's rank correlation was used to examine the correlation between two sets of data.

To identify the unique sequence variants of each strain, we identified variations by GATK v4.0.3 [110] and used the H1 haplotype of the Dd1 strain as the reference genome. The BEDTools was used to calculate the shared sequence variations between strains [125]. To calculate the divergence between Dd1 and other strains, the sequence of a diploid strain X was first compared with the reference genome (Dd1-H1). If a difference was detected, it would then be compared to another haplotype (Dd1-H2) of Dd1. Only if it differed from both Dd1-H1 and Dd1-H2, it was counted as a unique SNP (Additional file 2: Table S2). Such SNPs could be homozygous or heterozygous in the strain X, but it would be counted only once. The total SNP number was divided by the reference genome size to get the divergence (Additional file 2: Table S2). For identification of alleles derived from each parent in the DK1 and DK2 strains, we compared Dd1-H2, KM2, DK1, or DK2 sequences to the reference genome (Dd1-H1) and identified the variants. We then counted the variant numbers shared between different strains to draw the Venn diagram (Additional file 1: Fig. S8).

To calculate the evolutionary rates of individual genes, pairwise Ka/Ks was calculated for all alleles in the functional allele subgroup (Fig. 1b), and then all Ka/Ks values were averaged for each gene. Only the CDS subgroups containing more than three alleles were used in this analysis. The pairwise Ka/Ks value was calculated in PAML 4.9e [126].

### Quantitative PCR

The genomic DNA was diluted to an appropriate concentration and then subjected to quantitative PCR using gene-specific primers (Additional file 2: Table S10) and Fast SYBR Green master mix in an Applied Biosystems 7500 Fast Real-Time PCR System (Applied Biosystems,

Waltham, MA, USA). Data were analyzed using the built-in analysis program.

### Clustering analysis and the bootstrap method

We used the clusterdist function in ClusterScan v.0.2.2 [127] to calculate the number of genes that form clusters in the genome for the conserved group with the option `dist = 500`. We counted the number of conserved CN genes that were defined inside the clusters and used to calculate their proportion in the conserved CN group. To generate the reference genome background set, we randomly picked the same number of conserved CN genes from the whole genome 1000 times and calculated the number of genes in the clusters. After we acquired the distribution of clustering percentage from the random sampling datasets, we tested whether the observed clustering percentage of the conserved CN group was significantly deviated from the random distribution.

### Motif discovery

The aligned reads containing telomeric repeats on the left or right end with respect to the contig were mapped separately to the reference genome using BWA-MEM. We calculated the read depth of telomere-containing reads in a 0.5-kb interval using the SAMtools bedcov module. DNA sequences of the intervals having at least four telomere-containing reads were extracted and subjected to motif analysis after removing the telomeric sequences. Motif discovery was performed using MEME and the occurrence of motifs in the input regions was searched by FIMO in the MEME suite v5.0.5 [128].

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12915-020-00912-2>.

**Additional file 1: Figure S1.** De novo assembly of the *P. bursaria* MAC genome. **Figure S2.** *P. bursaria* introns have very conserved 5' and 3' splice sites. **Figure S3.** PFG Southern blot analysis of the genomic DNA of different *P. bursaria* strains. **Figure S4.** The extensive chromosome breakage pattern in the MAC. **Figure S5.** A model showing how highly variable breaking sites lead to non-uniform gene dosage. **Figure S6.** Contig copy number is uniform in *T. thermophila*. Different *P. bursaria* strains show similar patterns of copy number distribution. **Figure S7.** Conserved GA-rich motifs are found near chromosome breakage sites. **Figure S8.** DK1 and DK2 strains are the real F1 progeny of Dd1 and KM2.

**Additional file 2: Table S1.** Basic statistics for the coding regions of the two haplotypes of the Dd1 reference genome. **Table S2.** Statistics of sequence diversity for different *P. bursaria* strains compared to the H1 haplotype of Dd1. **Table S3.** Basic statistics for the intergenic and intron regions on two haplotypes of the reference genome. **Table S5.** Enriched GO terms for *P. bursaria* specific genes ( $p$ -value  $\leq 0.05$  after the Benjamini-Hochberg correction). **Table S7.** The correlation matrix (Spearman's correlation coefficient  $\rho$ ) of the gene copy number between genetically divergent strains. **Table S8.** Enriched GO terms for the conserved and non-conserved groups of genes ( $p$ -value  $\leq 0.05$  after the Benjamini-Hochberg correction). **Table S9.** DNA and RNA sequencing library list and data accession numbers. **Table S10.** Primer list used in qPCR and Southern blot and F1 progeny validation.

**Additional file 3: Table S4.** Splicing-related genes identified in the budding yeast *Saccharomyces cerevisiae*.

**Additional file 4: Table S6.** Read counts of Illumina and Pacbio reads with telomeric sequences at their ends.

### Abbreviations

CNV: Copy number variation; SNP: Single nucleotide polymorphism; MIC: Micronucleus; MAC: Macronucleus; IES: Internal eliminated sequence; NHEJ: Non-homologous end joining; CDS: Coding sequence; ISV: Internal structural variation region; CEG: Core eukaryotic gene; GO: Gene ontology; qPCR: Quantitative PCR; PBCV-1: *Paramecium bursaria* Chlorella virus-1

### Acknowledgements

We thank Meng-Chao Yao and members of the Leu lab for helpful discussion and comments on the manuscript. We also thank John O'Brien for manuscript editing and the IMB Genomics core and Academia Sinica Sequencing core for sequencing services.

### Authors' contributions

JYL conceived the study. YHC, CFJL, IJT, and JYL designed analyses and interpreted results. YHC, YTJ, and YHY performed the experiments. YHC, CFJL, and IJT performed the NGS data analysis. MF offered advice and technical assistance for carrying out the studies on *P. bursaria*. YHC and JYL wrote the paper. All authors read and approved the final manuscript.

### Funding

This work was supported by Academia Sinica of Taiwan (grant no. AS-IA-105-L01 and AS-TP-107-ML06 to JYL; AS-CDA-107-L01 to IJT) and the Taiwan Ministry of Science and Technology (MOST 107-2321-B-001-010 to JYL; 105-2628-B-001-002-MY3 to IJT).

### Availability of data and materials

The datasets generated and analyzed during the current study are included in this published article and its supplementary information files and available in NCBI under the accession number BioProject PRJNA556774 (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA556774>) and PRJNA555640 (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA555640>).

### Ethics approval and consent to participate

Not applicable

### Consent for publication

Not applicable

### Competing interests

The authors have declared that no competing interests exist.

### Author details

<sup>1</sup>Genome and Systems Biology Degree Program, Academia Sinica and National Taiwan University, Taipei 106, Taiwan. <sup>2</sup>Institute of Molecular Biology, Academia Sinica, 128 Sec. 2, Academia Road, Nankang, Taipei 115, Taiwan. <sup>3</sup>Biodiversity Research Center, Academia Sinica, Taipei 115, Taiwan. <sup>4</sup>Graduate School of Sciences and Technology for Innovation, Yamaguchi University, Yamaguchi 753-8512, Japan.

Received: 22 October 2019 Accepted: 29 October 2020

Published online: 30 November 2020

### References

- Olsen KM, Wendel JF. A bountiful harvest: genomic insights into crop domestication phenotypes. *Ann Rev Plant Biol.* 2013;64:47–70.
- Girirajan S, Campbell CD, Eichler EE. Human copy number variation and complex genetic disease. *Annu Rev Genet.* 2011;45:203–26.
- Lauer S, Gresham D. An evolving view of copy number variants. *Curr Genet.* 2019.
- Huang W, Massouras A, Inoue Y, Peiffer J, Ramia M, Tarone AM, et al. Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Res.* 2014;24(7):1193–208.
- Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. *Science.* 2004; 305(5683):525–8.



6. lafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. *Nat Genet.* 2004; 36(9):949–51.
7. Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet.* 2012; 44(7):803–7.
8. Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet.* 2011;43(10):956–63.
9. Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, et al. De novo rates and selection of large copy number variation. *Genome Res.* 2010; 20(11):1469–81.
10. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. *Nature.* 2010;464(7289):704–12.
11. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science.* 2007;315(5813):848–53.
12. Beckmann JS, Estivill X, Antonarakis SE. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat Rev Genet.* 2007;8(8):639–46.
13. Lye ZN, Purugganan MD. Copy number variation in domestication. *Trends Plant Sci.* 2019;24(4):352–65.
14. Dolatabadian A, Patel DA, Edwards D, Batley J. Copy number variation and disease resistance in plants. *TAG Theor Appl Genet.* 2017;130(12):2479–90.
15. Weetman D, Djogbenou LS, Lucas E. Copy number variation (CNV) and insecticide resistance in mosquitoes: evolving knowledge or an evolving problem? *Curr Opin Insect Sci.* 2018;27:82–8.
16. Farslow JC, Lipinski KJ, Packard LB, Edgley ML, Taylor J, Flibotte S, et al. Rapid increase in frequency of gene copy-number variants during experimental evolution in *Caenorhabditis elegans*. *BMC Genomics.* 2015;16: 1044.
17. Bellec L, Katz LA. Analyses of chromosome copy number and expression level of four genes in the ciliate *Chilodonella uncinata* reveal a complex pattern that suggests epigenetic regulation. *Gene.* 2012;504(2):303–8.
18. Xu K, Doak TG, Lipps HJ, Wang JM, Swart EC, Chang WJ. Copy number variations of 11 macronuclear chromosomes and their gene expression in *Oxytricha trifallax*. *Gene.* 2012;505(1):75–80.
19. Huang LJ, Lu XF, Zhu CY, Lin XF, Yi ZZ. Macronuclear Actin copy number variations in single cells of different *Pseudokeronopsis* (Alveolata, Ciliophora) populations. *Eur J Protistol.* 2017;59:75–81.
20. Heyse G, Jonsson F, Chang WJ, Lipps HJ. RNA-dependent control of gene amplification. *Proc Natl Acad Sci U S A.* 2010;107(51):22134–9.
21. Nowacki M, Haye JE, Fang W, Vijayan V, Landweber LF. RNA-mediated epigenetic regulation of DNA copy number. *Proc Natl Acad Sci U S A.* 2010; 107(51):22140–4.
22. Schoeberl UE, Mochizuki K. Keeping the soma free of transposons: programmed DNA elimination in ciliates. *J Biol Chem.* 2011;286(43):37045–52.
23. Yao MC. Modulating somatic DNA copy number through maternal RNA. *Proc Natl Acad Sci U S A.* 2010;107(51):21951–2.
24. Yao MC, Chao JL. RNA-guided DNA deletion in *Tetrahymena*: an RNAi-based mechanism for programmed genome rearrangements. *Annu Rev Genet.* 2005;39:537–59.
25. Yerlici VT, Landweber LF. Programmed genome rearrangements in the ciliate *Oxytricha*. *Microbiol Spectr.* 2014;2(6) <https://doi.org/10.1128/microbiolspec.MDNA3-0025-2014>.
26. Betermier M, Duhaucourt S. Programmed rearrangement in ciliates: *Paramecium*. *Microbiol Spectr.* 2014;2(6) <https://doi.org/10.1128/microbiolspec.MDNA3-0035-2014>.
27. Yao MC, Zheng KQ, Yao CH. A conserved nucleotide-sequence at the sites of developmentally regulated chromosomal breakage in *Tetrahymena*. *Cell.* 1987;48(5):779–88.
28. Challoner PB, Blackburn EH. Conservation of sequences adjacent to the telomeric C<sub>4</sub>A<sub>2</sub> repeats of ciliate macronuclear ribosomal-RNA gene molecules. *Nucleic Acids Res.* 1986;14(15):6299–311.
29. Yao MC, Yao CH, Monks B. The controlling sequence for site-specific chromosome breakage in *Tetrahymena*. *Cell.* 1990;63(4):763–72.
30. Fan Q, Yao M. New telomere formation coupled with site-specific chromosome breakage in *Tetrahymena thermophila*. *Mol Cell Biol.* 1996; 16(3):1267–74.
31. Eisen JA, Coyne RS, Wu M, Wu DY, Thiagarajan M, Wortman JR, et al. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.* 2006;4(9):1620–42.
32. Yao MC, Chao JL, Cheng CY. Programmed genome rearrangements in *Tetrahymena*. *Microbiol Spectr.* 2014;2(6) <https://doi.org/10.1128/microbiolspec.MDNA3-0012-2014>.
33. Hamilton EP, Kapusta A, Huvos PE, Bidwell SL, Zafar N, Tang H, et al. Structure of the germline genome of *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome. *Elife.* 2016;5.
34. Le Mouel A, Butler A, Caron F, Meyer E. Developmentally regulated chromosome fragmentation linked to imprecise elimination of repeated sequences in *paramecia*. *Eukaryot Cell.* 2003;2(5):1076–90.
35. Gilley D, Blackburn EH. Lack of telomere shortening during senescence in *Paramecium*. *Proc Natl Acad Sci U S A.* 1994;91(5):1955–8.
36. Swart EC, Bracht JR, Magrini V, Minx P, Chen X, Zhou Y, et al. The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. *PLoS Biol.* 2013;11(1):e1001473.
37. Chen X, Bracht JR, Goldman AD, Dolzhenko E, Clay DM, Swart EC, et al. The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell.* 2014;158(5):1187–98.
38. Prescott DM. The DNA of ciliated protozoa. *Microbiol Rev.* 1994;58(2):233–67.
39. Siegel RW, Karakashian SJ. Dissociation and restoration of endocellular symbiosis in *Paramecium-bursaria*. *Anat Rec.* 1959;134(3):639.
40. Kreutz M, Stoeck T, Foissner W. Morphological and molecular characterization of *Paramecium* (*Vrirdoparamecium* nov. subgen.) *chlorelligerum* Kahl (Ciliophora). *J Eukaryot Microbiol.* 2012;59(6):548–63.
41. Fokin SI, Przybos E, Chivilev SM, Beier CL, Horn M, Skotarczak B, et al. Morphological and molecular investigations of *Paramecium schewiakoffi* sp nov (Ciliophora, Oligohymenophorea) and current status of distribution and taxonomy of *Paramecium* spp. *Eur J Protistol.* 2004;40(3):225–43.
42. Kodama Y, Fujishima M. Cycloheximide induces synchronous swelling of perialgal vacuoles enclosing symbiotic *Chlorella vulgaris* and digestion of the algae in the ciliate *Paramecium bursaria*. *Protist.* 2008;159(3):483–94.
43. Karakashian SJ, Rudzinska MA. Inhibition of lysosomal fusion with symbiont-containing vacuoles in *Paramecium bursaria*. *Exp Cell Res.* 1981;131(2):387–93.
44. Gu FK, Chen L, Ni B, Zhang XM. A comparative study on the electron microscopic enzymo-cytochemistry of *Paramecium bursaria* from light and dark cultures. *Eur J Protistol.* 2002;38(3):267–78.
45. Hoshina R, Imamura N. Multiple origins of the symbioses in *Paramecium bursaria*. *Protist.* 2008;159(1):53–63.
46. McGrath CL, Gout JF, Doak TG, Yanagi A, Lynch M. Insights into three whole-genome duplications gleaned from the *Paramecium caudatum* genome sequence. *Genetics.* 2014;197(4):1417–28.
47. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, et al. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature.* 2006;444(7116):171–8.
48. Arnaiz O, Mathy N, Baudry C, Malinsky S, Aury JM, Denby Wilkes C, et al. The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. *PLoS Genet.* 2012;8(10):e1002984.
49. Coyne RS, Thiagarajan M, Jones KM, Wortman JR, Tallon LJ, Haas BJ, et al. Refined annotation and assembly of the *Tetrahymena thermophila* genome sequence through EST analysis, comparative genomic hybridization, and targeted gap closure. *BMC Genomics.* 2008;9.
50. Rautian MS, Potekhin AA. Elektrokaryotypes of macronuclei of several *Paramecium* species. *J Eukaryot Microbiol.* 2002;49(4):296–304.
51. Kraut H, Lipps HJ, Prescott DM. The genome of hypotrichous ciliates. *Int Rev Cytol.* 1986;99:1–28.
52. Cullis CA. DNA amounts in the nuclei of *Paramecium bursaria*. *Chromosoma.* 1973;40(2):127–33.
53. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature.* 2015; 526(7571):68–74.
54. Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, et al. The diploid genome sequence of *Candida albicans*. *Proc Natl Acad Sci U S A.* 2004;101(19):7329–34.
55. Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, et al. The *Drosophila* genome nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. *Genetics.* 2015;199(4):1229–41.



56. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*. 2007;23(9):1061–7.
57. Jaillon O, Bouhouche K, Gout JF, Aury JM, Noel B, Saudemont B, et al. Translational control of intron splicing in eukaryotes. *Nature*. 2008;451(7176):359–62.
58. Bergman CM, Pfeiffer BD, Rincon-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, et al. Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol*. 2002;3(12):RESEARCH0086.
59. Tatarinova TV, Chekalin E, Nikolsky Y, Bruskin S, Chebotarov D, McNally KL, et al. Nucleotide diversity analysis highlights functionally important genomic regions. *Sci Rep*. 2016;6:35730.
60. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res*. 2004;14(6):1188–90.
61. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 2015;16.
62. McMahon HT, Mills IG. COP and clathrin-coated vesicle budding: different pathways, common approaches. *Curr Opin Cell Biol*. 2004;16(4):379–91.
63. Rout MP, Field MC. The evolution of organellar coat complexes and organization of the eukaryotic cell. *Annu Rev Biochem*. 2017;86:637–57.
64. Karakashian MW, Karakashian SJ. Intracellular digestion and symbiosis in *Paramecium bursaria*. *Exp Cell Res*. 1973;81(1):111–9.
65. Kodama Y, Fujishima M. Secondary symbiosis between *Paramecium* and *Chlorella* cells. *Int Rev Cell Mol Biol*. 2010;279:33–77.
66. Zhang L, Lee SY, Beznoussenko GV, Peters PJ, Yang JS, Gilbert HY, et al. A role for the host coatomer and KDEL receptor in early vaccinia biogenesis. *Proc Natl Acad Sci U S A*. 2009;106(1):163–8.
67. Panchy N, Lehti-Shiu M, Shiu SH. Evolution of gene duplication in plants. *Plant Physiol*. 2016;171(4):2294–316.
68. Kikuchi T, Eves-van den Akker S, Jones JT. Genome evolution of plant-parasitic nematodes. *Annu Rev Phytopathol*. 2017;55:333–54.
69. Yano K, Shibata S, Chen WL, Sato S, Kaneko T, Jurkiewicz A, et al. CERBERUS, a novel U-box protein containing WD-40 repeats, is required for formation of the infection thread and nodule development in the legume-Rhizobium symbiosis. *Plant J*. 2009;60(1):168–80.
70. Finelli MJ, Oliver PL. TLDc proteins: new players in the oxidative stress response and neurological disease. *Mamm Genome*. 2017;28(9–10):395–406.
71. Durand M, Kolpak A, Farrell T, Elliott NA, Shao W, Brown M, et al. The OXR domain defines a conserved family of eukaryotic oxidation resistance proteins. *BMC Cell Biol*. 2007;8:13.
72. Dykens JA, Shick JM, Benoit C, Buettner GR, Winston GW. Oxygen radical production in the sea-anemone *Anthopleura elegantissima* and its endosymbiotic algae. *J Exp Biol*. 1992;168:219–41.
73. Finger I, Audi D, Bernstein M, Voremberg S, Harkins K, Birnbaum M, et al. Switching of *Paramecium* surface antigen types with purified antigen and conditioned medium containing 70 kD proteins. *Archiv Fur Protistenkunde*. 1996;146(3–4):373–81.
74. Matsuda A, Forney JD. Analysis of *Paramecium tetraurelia* A-51 surface antigen gene mutants reveals positive-feedback mechanisms for maintenance of expression and temperature-induced activation. *Eukaryot Cell*. 2005;4(10):1613–9.
75. Gratiás A, Betermier M. Processing of double-strand breaks is involved in the precise excision of *Paramecium* internal eliminated sequences. *Mol Cell Biol*. 2003;23(20):7152–62.
76. Pritchard AE, Cummings DJ. Structural and functional analysis of the origin of replication of mitochondrial DNA from *Paramecium aurelia*: I. Inverted complements form the terminal loop. *Curr Genet*. 1984;8(7):477–82.
77. Pritchard AE, Herron LM, Cummings DJ. Cloning and characterization of *Paramecium* mitochondrial DNA replication initiation regions. *Gene*. 1980;11(1–2):43–52.
78. Baroin A, Prat A, Caron F. Telomeric site position heterogeneity in macronuclear DNA of *Paramecium primaurelia*. *Nucleic Acids Res*. 1987;15(4):1717–28.
79. Keller AM, Le Mouel A, Caron F, Katinka M, Meyer E. The differential expression of the G surface antigen alleles in *Paramecium primaurelia* heterozygous cells correlates to macronuclear DNA rearrangement. *Dev Genet*. 1992;13(4):306–17.
80. Forney JD, Blackburn EH. Developmentally controlled telomere addition in wild-type and mutant *paramecia*. *Mol Cell Biol*. 1988;8(1):251–8.
81. Duret L, Cohen J, Jubin C, Dessen P, Gout JF, Mousset S, et al. Analysis of sequence variability in the macronuclear DNA of *Paramecium tetraurelia*: a somatic view of the germline. *Genome Res*. 2008;18(4):585–96.
82. La Terza A, Miceli C, Luporini P. Differential amplification of pheromone genes of the ciliate *Euplotes raikovi*. *Dev Genet*. 1995;17(3):272–9.
83. Hirsh AE, Fraser HB. Protein dispensability and rate of evolution. *Nature*. 2001;411(6841):1046–9.
84. Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res*. 2002;12(6):962–8.
85. Zhang J, Yang JR. Determinants of the rate of protein sequence evolution. *Nat Rev Genet*. 2015;16(7):409–20.
86. Pal C, Papp B, Hurst LD. Highly expressed genes in yeast evolve slowly. *Genetics*. 2001;158(2):927–31.
87. Yanagi A, Haga N. Induction of conjugation by methyl cellulose in *Paramecium*. *J Eukaryot Microbiol*. 1998;45(1):87–90.
88. He M, Wang JF, Fan XP, Liu XH, Shi WY, Huang N, et al. Genetic basis for the establishment of endosymbiosis in *Paramecium*. *ISME J*. 2019;13(5):1360–9.
89. Nelson CE, Hersh BM, Carroll SB. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol*. 2004;5(4):R25.
90. Hurowitz EH, Brown PO. Genome-wide analysis of mRNA lengths in *Saccharomyces cerevisiae*. *Genome Biol*. 2003;5(1):R2.
91. Hudson AJ, Stark MR, Fast NM, Russell AG, Rader SD. Splicing diversity revealed by reduced spliceosomes in *C. merolae* and other organisms. *RNA Biol*. 2015;12(11):1–8.
92. Austerberry CF, Allis CD, Yao MC. Specific DNA rearrangements in synchronously developing nuclei of *Tetrahymena*. *P Natl Acad Sci-Biol*. 1984;81(23):7383–7.
93. Chen X, Jung S, Beh LY, Eddy SR, Landweber LF. Combinatorial DNA rearrangement facilitates the origin of new genes in ciliates. *Genome Biol Evol*. 2015;7(10):2859–70.
94. Qi J, Chen Y, Copenhaver GP, Ma H. Detection of genomic variations and DNA polymorphisms and impact on analysis of meiotic recombination and genetic mapping. *Proc Natl Acad Sci U S A*. 2014;111(27):10007–12.
95. Zhou Y, Wubneh H, Schwarz C, Landweber LF. A chimeric chromosome in the ciliate *oxytricha* resulting from duplication. *J Mol Evol*. 2011;73(3–4):70–3.
96. Chen X, Jiang Y, Gao F, Zheng W, Krock TJ, Stover NA, et al. Genome analyses of the new model protist *Euplotes vannus* focusing on genome rearrangement and resistance to environmental stressors. *Mol Ecol Resour*. 2019;19(5):1292–308.
97. Klobutcher LA. Characterization of in vivo developmental chromosome fragmentation intermediates in *E. crassus*. *Mol Cell*. 1999;4(5):695–704.
98. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics*. 2011;27(7):1017–8.
99. Schlattl A, Anders S, Waszak SM, Huber W, Korbel JO. Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. *Genome Res*. 2011;21(12):2004–13.
100. Selmecki AM, Dulmage K, Cowen LE, Anderson JB, Berman J. Acquisition of aneuploidy provides increased fitness during the evolution of antifungal drug resistance. *PLoS Genet*. 2009;5(10):e1000705.
101. Zhu J, Tsai HJ, Gordon MR, Li R. Cellular stress associated with aneuploidy. *Dev Cell*. 2018;44(4):420–31.
102. Kodama Y, Inouye I, Fujishima M. Symbiotic *Chlorella vulgaris* of the ciliate *Paramecium bursaria* plays an important role in maintaining perialgal vacuole membrane functions. *Protist*. 2011;162(2):288–303.
103. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017;14(4):417–9.
104. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*. 2016;13(12):1050–4.
105. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 2014;9(11):e12963.
106. Pryszcz LP, Nemeth T, Gacser A, Gabaldon T. Genome comparison of *Candida orthopsilosis* clinical strains reveals the existence of hybrids between two distinct subspecies. *Genome Biol Evol*. 2014;6(5):1069–78.
107. Vinson JP, Jaffe DB, O'Neill K, Karlsson EK, Stange-Thomann N, Anderson S, et al. Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. *Genome Res*. 2005;15(8):1127–35.

108. Prysacz LP, Gabaldon T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* 2016;44(12):e1113.
109. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. *Genome Biol.* 2004;5(2):R12.
110. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297–303.
111. Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* 2017;27(5):801–12.
112. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin).* 2012;6(2):80–92.
113. Arnaiz O, Van Dijk E, Betermier M, Lhuillier-Akakpo M, de Vanssay A, Duharcourt S, et al. Improved methods and resources for paramecium genomics: transcription units, gene annotation and gene expression. *BMC Genomics.* 2017;18(1):483.
114. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2010;26(5):589–95.
115. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997v2 [q-bioGN]. 2013.
116. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. *Nucleic Acids Res.* 2005;33(Web Server issue):W116–20.
117. Johri P, Krenek S, Marinov GK, Doak TG, Berendonk TU, Lynch M. Population genomics of *Paramecium* species. *Mol Biol Evol.* 2017;34(5):1194–216.
118. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32(5):1792–7.
119. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 2000;17(4):540–52.
120. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 2015;32(1):268–74.
121. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods.* 2017;14(6):587–9.
122. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol.* 2016;33(6):1635–8.
123. Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met.* 1995;57(1):289–300.
124. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25(16):2078–9.
125. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26(6):841–2.
126. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 2007;24(8):1586–91.
127. Volpe M, Miralto M, Gustincich S, Sanges R. ClusterScan: simple and generalistic identification of genomic clusters. *Bioinformatics.* 2018;34(22):3921–3.
128. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 2009;37(Web Server issue):W202–8.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

