Supplementary data

References

Supplementary Figures

Supplementary Tables

1 Choice of strain and DNA extraction

Paramecium, like all ciliates, possesses two kinds of nuclei, germinal micronuclei and a somatic macronucleus, the genome of which derives from the germinal one by DNA rearrangements during sexual processes, conjugation and autogamy. The rearrangements consist (1) in precise elimination of short sequences called internally eliminated sequences (IES) (2) imprecise elimination of transposons and other repeated sequences, usually leading to chromosome fragmentation (3) repair of chromosome ends by *de novo* telomere addition (4) amplification to high copy number (800n). We decided to sequence the macronuclear genome principally because it is stripped of repeated elements and of IESs, which are almost impossible to detect in the germinal sequence without knowing the rearranged version of the genome. Since the macronuclear DNA evolves with age by chromosome breakage and healing, we used a culture of relatively young cells, ~10 fissions since the previous autogamy.

Macronuclear DNA used for the shotgun libraries was of the same batch as the one used by[1] and prepared as described[2]. The strain used was stock d4-2, a derivative of stock 51 of *Paramecium tetraurelia*[3]. A culture of 30,000 autogamous cells of mating type VIII was grown for ten fissions in a final volume of ten liters of Wheat Grass Powder (Pines International, Lawrence Kansas) medium. Stationary cells, 100% non autogamous because autogamy is inhibited at the age of ten fissions, were harvested by centrifugation, washed twice and stored for two hours in 0.1 M Sodium Citrate, 0.1 M NaH_2PO_4, 0.1 M Na_2HPO_4, 0.1 M $CaCl_2$. The 1.5 ml cell pellet was then washed twice in cold 0.25 M sucrose, 2 mM $MgCl_2$ and then homogenized in 1 volume of 0.25 M sucrose, 2 mM $MgCl_2$, 10 mM Tris Cl pH 7.2 and 0.5 % Nonidet P-40. The lysate was made 15 ml in the same buffer and centrifuged twice 1 min at 100 g. The supernatant containing most of the mitochondria and the micronuclei (the sole sources of non macronuclear DNA[4]) was discarded. The final pellet was extracted by overnight lysis of the pellet in 25 ml of 0.5 M EDTA pH 9.0, 1 % sarkosyl, 1 % SDS, and 1 mg/ml proteinase K at 55°C. The macronuclear DNA was then purified by two phenol-chloroform extractions and one chloroform extraction, followed by centrifugation on a CsCl gradient.

2 Genome sequencing

The *P. tetraurelia* genome was sequenced using a whole genome shotgun approach. A total of 1,235,002 reads were generated from both ends of plasmid and BAC inserts (Table S1). The traces were assembled using Arachne[5], after elimination of the sequences matching the ribosomal DNA repeats, and the mitochondrial DNA. The version retained for further analysis contains the 697 supercontigs of >2 kb, with 861 sequence gaps.

3 cDNA sequencing

3.1 *P. tetraurelia* strains and cultivation

Strains d4-2 and 51 are entirely homozygous laboratory strains. d4-2 is a hybrid strain carrying a few genes from strain 29 in the genetic background of strain 51. Cells were grown in a wheat grass powder (Pines International Co., Lawrence, USA) infusion medium bacterized the day before use with *Klebsiella pneumoniae*, and supplemented with 0.8 mg/l of beta-sitosterol (Merck, Darmstadt, FRG). Unless otherwise stated, cells were maintained at 27°C. For conjugation and autogamy, the progression of cells through the different stages of nuclear reorganization was monitored by DAPI staining.

3.2 RNA purification

Total RNA was extracted from unwashed cell pellets using the TRIzol (Invitrogen) procedure, modified by the addition of glass beads. After the Trizol/Chloroform treatment, the supernatant was precipitated with isopropanol, the pellet was washed with ethanol and finally resuspended in DEPC-treated water. The RNA concentration was estimated by optical density at 260 and 280 nm.

3.3 Library construction

The cDNA libraries were constructed using the Cloneminer cDNA library construction Kit (Invitrogen) following the supplier’s instructions. Six differentiation stages were used to maximize the detection of genes expressed with a stage-enriched profile.

Library LK0AAA: Vegetative cells at high temperature (35°C). The temperature of the culture was raised from 27°C to 35°C in two hours, and then the culture was maintained during one hour before cell harvesting. The mean number of divisions per cell was 6 at that time. No sign of macronuclear fragmentation was observed after DAPI staining.

Library LK0ABA: Vegetative cells at high temperature (39°C). The temperature of the culture was raised from 27°C to 39°C in two hours and forty minutes, and then the culture was maintained during 35 minutes before cell harvesting. The mean number of divisions per cell was 6 at that time. Again, no sign of macronuclear fragmentation was observed after DAPI staining.

Library LK0ACA: Vegetative cells at the standard temperature (27°C).

Library LK0ADA: Conjugation (beginning of meiosis). The cells were conjugated after 3-4 divisions. The RNA was extracted 2 hours after the mixing of the two sexual types. 56% of the cells were engaged in conjugation, while the remaining 44% were sexually reactive but not engaged in meiosis.

Library LK0AEA: Conjugation (meiosis and beginning of macronuclear development). The cells were conjugated after 3-4 divisions. RNA was extracted 4.5 hours and 7 hours after mixing of the two sexual types. At t=4.5 hours, 45% of the cells were in meiosis, 11% were at the beginning of the fragmentation of the old macronucleus, and 44% were still not engaged in conjugation. At t=7h, 3% of the cells were in meiosis, 9% were at the beginning of the fragmentation of the old macronucleus, 44% had completed the fragmentation of the old macronucleus and entered the development of the new macronucleus, and 44% were still not engaged in conjugation. An equimolar amount of each extraction was mixed for library construction.

Library LK0AFA: Autogamy (macronuclear development). RNA was extracted 11 hours and 20 hours after an arbitrarily chosen time point where 56% of the cells had started the fragmentation of the old macronucleus. At t=11 hours, 57% of the cells were at the beginning of the development of the new macronucleus, while 41% had a clearly visible new macronucleus, the remainder being at an early stage of fragmentation. At t= 20 hours, 8% of the cells had begun the development of the new macronucleus, while 89% had a clearly visible new macronucleus, the remainder being at an early stage of fragmentation.

3.4 Sequencing statistics

Sequencing was performed from the 5’ end of the insert. Statistics on the number of valid reads per library are presented in Table S4.

3.5 Estimation of the completion of the assembly

We used the EST data as an independent resource to further estimate the completeness of the genome assembly. We matcedh the sequences of 85,711 cDNA clones that have a significant match with the 13x shotgun reads (criteria used for BLAST alignments W=20, X=8) on the 697 scaffolds to see how many genes are part of regions excluded during the assembly process. All the clones that do not match the scaffolds were ribosomal RNA and mitochondrial contaminants, thus indicating that all genes covered by an EST were at least partly represented in the assembly.

4 Genome Annotation

4.1 Annotation Procedure

4.1.1 Repeat Masking

Most of the genome comparisons were performed with repeat masked sequences. For this purpose, we searched and masked sequentially several kinds of repeats:

- known *Paramecium* telomeric satellite repeats
- other known repeats and transposons available in Repbase with the Repeat masker program[6]
- Tandem repeats with the TRF program[7]

4.1.2 Exofish comparisons

Exofish[8] comparisons were performed at Genoscope, on a cluster of 40 CPU alpha EV6.8, with the Biofacet software package from Gene-IT (www.gene-it.com). When ecores (evolutionarily conserved regions) were contiguous in the two genomes, they were included in the same ecotig[9] (contig of ecores).

4.1.3 Exofish between the *Plasmodium falciparum* genome and *Paramecium*

Exofish comparison between *Plasmodium* and *Paramecium* was performed using TBLASTX with the following parameters: W=5, X=8, T=50, matches=10 and mismatches=-1. HSPs were filtered according to their length and percent identity.

4.1.4 Auto-Exofish between the *Paramecium* genome and itself

We also used Exofish to compare the *Paramecium* genome to itself since duplication events appeared to be old enough to allow a good degree of divergence between under-selected regions. Comparison was performed using TBLASTX with the following parameters: W=5, X=8, T=50, match=10, mismatches=-1. As introns are very short, stop codons were highly penalised (-90) to avoid letting HSPs cross over them. HSPs were filtered according to their length and percent identity.

4.1.5 Genewise

The Uniprot[10] database was used to detect well conserved genes between *Paramecium* and other species. As Genewise[11] is time greedy, the Uniprot database was first aligned with the *Paramecium* genome assembly using BLASTX (with parameters W=4, T=44 and a evalue cutoff at 10^{-2}). HSPs from the same protein were clustered on the genomic position, to assign one (or several) loci to each peptide. For a given locus, the five best matches were chosen for a Genewise alignment. A non-redundant database of ciliate peptides was aligned with the *Paramecium* genome assembly using the same pipeline.

4.1.6 Geneid, SNAP and GlimmerHMM

Geneid[12], SNAP[13] and GlimmerHMM[14] *ab inito* gene prediction software were trained on 247 *Paramecium* genes that had been annotated and reviewed by human experts.

4.1.7 SGP2

The SGP2 *ab initio* annotation software was trained according to[15].

4.1.8 *Paramecium* cDNAs

4.1.8.1 Alignment of cDNA sequences to the genomic reference sequence

A two-step strategy was used to align the *Paramecium tetraurelia* cDNA clones on the genomic reference sequence[16,17]. Preliminary transcript models were created based on the alignments of the 5' and 3' repeat-masked EST sequence reads derived from the cDNA clones and the *Paramecium tetraurelia* genome assembly. The repeats taken into account by the masking procedure were limited to microsatellites. The HSPs obtained by the BLAST[18] comparisons (W=20, X=8) were combined in a coherent manner, consistent with their position on the reference genomic sequence. In this way, one or several models were built for each transcript, composed of one or several tentative exons based on the alignment with the genome sequence. The model with the highest total score defined by the sum of the scores of each HSP (total score = 800) was selected as the preliminary transcript model that underwent further analysis. cDNA clones with discrepant alignments of their 5' and 3' sequences on the genome were considered to be putative chimeras and were excluded from the analysis.

The unmasked regions of such preliminary transcript models were extended by 5 kb of genomic sequence on each end, and realigned with the cDNA clones using the est2genome[19] algorithm (-mismatch 3 -gap_penalty 6 -align 1 -space 500). This procedure defined transcript models with a high fraction of *bona fide* intron-exon boundaries. Such transcript models, supported by the cDNA clones, were obtained for 90% of the cDNA resource.

These transcript models were fused in gene models by a single linkage clustering approach, in which transcript models from the same genomic region and same strand sharing at least 100 bp are merged in a single model.

4.1.8.2 Mapping of the cDNA clones to the paralogous genes in *Paramecium tetraurelia*

The same two-step strategy was used to align the cDNA clones to the paralogous genes. The 5’ and 3’ repeat-masked sequence reads from each cDNA clone were aligned with the genome assembly using BLAST, this time lowering the stringency when building the preliminary transcript models. In this way, all models with a total score corresponding to 50% of the “best match” were kept as preliminary transcript models. The exon boundaries and the structure of each of the transcript models were further defined using the est2genome algorithm, using -mismatch=2:-gap_penalty=4:-align=1:-space=500 as parameters.

4.1.9 Alveolata mRNAs

A collection of 129,145 public mRNAs (from the Alveolata clade) was first aligned with the *Paramecium* genome assembly using Blat[20]. This database was composed of public mRNAs downloaded from the NCBI[21] and clusters of ESTs from the TIGR gene indices[22]. To refine Blat alignment, we used Est2Genome[19]. The best match of each mRNA was chosen for an alignment with Est2Genome. Matches with a score near the best score (maximum 20% of the best score) were also aligned. Blat alignments were made using default parameters between translated genomic and translated mRNAs with a maximum intron size of 150 nucleotides. Est2Genome was used with the following parameters: match=1, mismatch=-1 and gap penalty=3.

4.1.10 Remapped annotations

1,164 genes that had been annotated and reviewed by human experts were remapped using est2genome on the genome assembly of *Paramecium*. Only perfect alignments (from start to stop with correct splice sites), a total of 1,023 genes, were integrated with other predictions.

4.1.11 Integration of resources using GAZE

All the resources described here were used to automatically build *Paramecium* gene models using GAZE[23]. Individual predictions from each of the programs (geneid, SNAP, glimmerHMM, SGP2, exofish, genewise and est2genome) were broken down into segments (coding, intron, intergenic) and signals (start codon, stop codon, splice acceptor, splice donor, transcript start, transcript stop).

Exons predicted by ab-initio software, exofish, genewise, and est2genome were used as coding segments. Introns predicted by genewise and est2Genome were used as intron segments. Intergenic segments created from the span of each mRNA, with a negative score (coercing GAZE not to split genes). Predicted repeats were used as intron segments, and non-coding RNAs as intergenic segments, to avoid prediction of genes coding proteins in such regions.

The whole genome was scanned to find signals (splice sites, start and stop codons), and two signals, transcript start and stop, were extracted from the ends of mRNAs.

Each segment extracted from a software output which predicts exon boundaries (like genewise, est2genome or ab-initio predictors), was used by GAZE only if GAZE chose the same boundaries. Each segment or signal from a given program was given a value reflecting our confidence in the data, and these values were used as scores for the arcs of the GAZE automaton (Fig S2). All signals were given a fixed score, but segment scores were context sensitive: coding segment scores were linked to the percentage identity (%ID) of the alignment; intronic segment scores were linked to the %ID of the flanking exons. The impact of each data source (exofish, geneid, etc.) was evaluated on a reference sequence (see 4.2), and a weight was assigned to each resource to further reflect its reliability and accuracy in predicting gene models. This weight acts as a multiplier for the score of each information source, before processing by GAZE.

When applied to the entire assembled sequence, GAZE predicts 39,642 gene models.

4.2 Evaluation of the *Paramecium* gene annotation

4.2.1 Use of reference sequences to evaluate gene annotation accuracy

Each prediction method was evaluated on a reference *Paramecium* sequence. To be closest to the annotation conditions, we used the assembly version of a finished sequence[24] (the longest chromosome, named megabase) containing about 450 manually annotated genes. Gene annotations of the megabase were remapped on our assembly version (named scaffold_1) using est2genome[19] (i.e. the virtual mRNAs of manually annotated genes were aligned on scaffold_1). Of the 464 genes extracted from the megabase, 412 were remapped perfectly (100% of the mRNA with 100% identity), 33 alignments containing gaps (which may result in frameshifts in the coding sequence), 15 alignments containing mismatches and 4 resulting in a bad structure (without start, stop or splice sites). So we evaluated the accuracy of each prediction method on 412 genes from scaffold_1 (Fig S3, S4 and S5).

4.2.2 Global statistics for the whole annotation

GAZE predicted 39,642 gene models on the whole genome, corresponding to a high coding density (about 77% of the entire assembly, compared to 1% of the Human genome, Fig S6). The average number of exons per gene is about 3.3, lower than in mammals, but the average protein size is similar (about 450 amino acids, Fig S7), involving a higher average size of coding exons. The intron size distribution is very atypical, 20 - 30 nucleotides.

Of the 39,642 annotated gene models, a large fraction is supported by at least one biological evidence. Only 763 gene models (about 2%) are only supported by ab-initio predictions (Table S5). About 8,000 genes are only annotated using *Paramecium* specific resources (cDNAs, exofish or ab-initio software calibrated using *Paramecium* genes), which is not surprising given the very large evolutionary distance to even the closest sequenced genomes.

5 Identification of orthologous genes

5.1 Proteome comparison between representative species

We identified orthologous genes among 45 pairs of genomes from ten species: human, *Arabidopsis thaliana*, *Paramecium tetraurelia*, *Plasmodium falciparum, Tetrahymena thermophila, Drosophila melanogaster, Neurospora crassa, Dictyostelium discoideum, Thalassiosira pseudonana* and *Cyanidioschyzon merolae*. Each pair of predicted gene sets was aligned with the Smith-Waterman algorithm, and alignments with a score higher than 300 (BLOSUM62, gapo=10, gape=1) were retained (Fig S8 and Table S6). Two genes, A from genome GA and B from genome GB, were considered orthologs if B is the best match of gene A in GB and A is the best match of B in GA.

Predicted peptides for each genome were:

Human:	22,218 peptides at Ensembl (version 31.35d)
Drosophila:	13,792 peptides at Ensembl (version 29.3d)
Paramecium:	39,642 peptides at Genoscope (this report)
Plasmodium:	5,365 peptides at PlasmoDB (released november 2004)
Tetrahymena:	27,430 peptides at TIGR (released august 2004)
Arabidopsis:	26,639 peptides at MIPS (released february 2004)
Neurospora:	10,620 peptides at Broad Institute (released February 2005)
Dictyostelium:	13,573 peptides at dictyBase (released may 2005)
Thalassiosira:	11,397 peptides at JGI (release 1.0)
Cyanidioschyzon:	5,013 peptides at MerolaeBase (released april 2004)

5.2 Pfam domain comparisons

Pfam domains[25] were used to study the distribution of *Paramecium* domains among domains shared by *Arabidopsis thaliana* and *Homo sapiens*. This distribution was compared to *Dictyostelium discoideum* and *Neurospora crassa* distributions. First of all, the common domain set and the exclusive domain set between *H. sapiens* and *A. thaliana* were established from InterPro-Scan results. Then, domains of each set were tagged present or absent in *Paramecium*. This last step was repeated with *D. discoideum* and *N. crassa* domains (Fig S9).

5.3 Search for genes with potential red algal origin

5.3.1 Detection using Pfam domains

Pfam domains[25] were used to select domains specific to plants, red algae and *Paramecium*. Then a list of 4038 Pfam domains present in *Paramecium*, *Cyanidoschyzon merolae, Homo sapiens*, *Arabidopsis thaliana*, *Dictyostelium discoideum*, *Neurospora crassa*, *Plasmodium falciparum* and *Thalassiosira pseudonana* was assembled. For each of these species, a flag indicates the presence or absence of each domain.
Comparing *A. thaliana*, *H. sapiens* and *N. crassa* Pfam domains, 522 Pfam domains were tagged as *A. thaliana* specific. Among those 522 Pfam domains, 27 were found in *Paramecium*, 33 in *P. falciparum*, 146 in *T. pseudonana* and 174 in *C. merolae*. Besides, among the 27 Pfam domains found in *Paramecium*, there are 20 Pfam domains which are present in *C. merolae*. When looking at the distribution of these 20 domains, none are restricted to plants and red algae, but also occur in other metazoans or Fungi than those tested initially, or in protists.

5.3.2 Detection by orthologous matches

Paramecium proteins with orthologs in *Homo sapiens*, *Arabidopsis thaliana* or *Cyanidoschyzon merolae* were extracted. From this set of proteins, 298 *Paramecium* proteins have orthologous proteins in *C. merolae* and *A. thaliana* only. 66 of those proteins have orthologous proteins in *C. merolae*, and 19 of them were aligned on 80% at least of their length. These 19 proteins show no restricted distribution to plants, red algae and *Paramecium*, when their sequences were blasted against Uniprot. Therefore, we failed to detect any protein specific to the red algae and plant lineages present in the *Paramecium* proteome.

5.4 Detection of chloroplast sequences

To search for paramecium genes that might be of chloroplastic origin we extracted from Uniprot and GenBank two sets of proteins, called respectively the 'ingroup' and the 'outgroup'. The ingroup contains proteins encoded by plastids of alveolate species (176 sequences), cyanobacterial proteins (24,794 proteins from 8 complete genomes), and proteins encoded by chloroplastic genomes (34,414 proteins). The outgroup contains protein sequences from 173 complete genomes not related to chloroplastic lineages (5 animals, 1 fungi, 16 archaea and 151 non-cyanobacterial bacteria). We then compared all *P. tetraurelia* proteins against both datasets with BLASTP (with a threshold E-value of 10^{-8}). We found 20 *P. tetraurelia* proteins (or protein families) for which the best BLASTP score was with an ingroup protein. For each of these 20 candidates we computed a multiple alignment and phylogenetic tree including all *P. tetraurelia* proteins and their ingroup and outgroup homologs. We found only three gene families for which the phylogenetic tree indicated a closer relationship of *P. tetraurelia* to cyanobacteria than to outgroup species. However, none of these genes is homologous to a known chloroplastic gene. We therefore cannot exclude that they were acquired by independent horizontal transfer events, unrelated to the endosymbiosis of the chloroplast-containing organism.

6 Identification of paralogous genes and construction of paralogons

Initially an all-against-all comparison of *Paramecium* predicted proteins was performed using Smith-Waterman algorithm and alignments with an evalue lower than 0.1 were retained.

The first step was to pair scaffold regions. Each scaffold was scanned with a sliding window where Best Reciprocal Hits (BRH) were used to pair genes. We associated each window of size w with a target scaffold, if at least p percent of the w genes matched the same scaffold. Duplicated regions (called paralogous blocks) were obtained using this windowing strategy, by merging contiguous windows associated to a common target scaffold.

The next step led to a bijective relation between paralogous blocks, defined during the first step, in order to obtain paralogons. We have defined a paralogon as a pair of paralogous blocks that could be recognized as deriving from a common ancestral region. The target of a given paralogous block is a scaffold region, which could be split into multiple blocks. To define paralogons, we need to associate paralogous blocks in a bijective manner. Associations of paralogous blocks were scored using the number of genes contained, according to the following rules. One gene A was considered if the region containing its paralog B is associated with the current block, and if A and B were not already put in any other paralogon. We treated each paralogon ranked by score; each gene from both paralogous blocks was assigned to its paralogon.

In a third step, from the set of paralogons, we attempted to increase the total number of duplicated genes. For that purpose, we tried, for each single gene (gene without paralog, and thus without BRH) to find a syntenic match among its ten best matches. On the complete *Paramecium* gene catalogue, we found 131 ancestral regions, covering about 90% of the initial gene set, and the third step allowed us to add 1,477 paralogous relations.

As many paralogy relationships remained unexplained by the recent Whole Genome Duplication (Fig S10), paralogon construction was launched again, from these 131 ancestral regions, using ancestral regions as sequences, with arbitrary ancestral gene order along this last one. The procedure was iterated a third time and showed 4 different events of whole genome duplication (Table S7). Parameters were:

- 1st round : w = 10 and p=60%
- 2nd round : w=10 and p=41%
- 3rd round and 4th round : w=20 and p=30%

Ks and Ka values were calculated on the entire set of paralogs for each WGD, using the codeml software from PAML package[26]. We define Ks as the number of silent or synonymous nucleotidic substitutions per site (mutations that do not change the amino acid), and Ka as the number of non-synonymous substitutions per site (mutations that change the amino acid). Distribution of Ks, Ka and Ka/Ks are shown in Figure S16, S17 and S18, and distribution of paralogous families are shown in Figure S19.

7 Remapping of metabolic pathways

7.1 Enzyme annotation

Enzyme detection in predicted *Paramecium* proteins was performed with PRIAM[27], using the PRIAM July 2004 ENZYME release.

864 different EC numbers, corresponding to enzyme domains, are associated with 5,617 *Paramecium* proteins. Therefore, about 14% of *Paramecium* proteins contain at least one enzymatic domain.

7.2 Association of metabolic pathways with enzymes and *Paramecium* proteins

From EC numbers, potential metabolic pathways are deduced using the KEGG pathway database[28]. Links between EC numbers and metabolic pathways were obtained from the KEGG website. Using this file and the PRIAM results, the 5617 *Paramecium* proteins which have an EC number were assigned to 119 pathways.

Following the KEGG pathway hierarchy, pathways from the same family were grouped together. For instance, glycolysis and TCA cycle belong to Carbohydrate metabolism. In this way, the different pathways found in *Paramecium* define 16 pathway families.

7.3 Metabolic pathways and evolution of copy number of enzyme coding-genes through successive WGDs

To study evolution of gene copy number involved in metabolic pathways through successive WGDs, two sets of genes were created for each WGD: one containing duplicated genes which have no EC numbers already assigned to a single gene (SPEC2x), and a second one containing single genes which have no EC numbers already assigned to a duplicated gene (SPEC1x).

For each pathway family, genes belonging to SPEC2x and SPEC1x were enumerated, and a ratio R was computed as follows:

$$R = \frac{N_{specific}}{N_{all}}$$

with, $N_{specific} = \dfrac{\text{Number of duplicated genes corresponding to EC numbers 2x specific}}{\text{Number of single genes corresponding to EC numbers 1x specific}}$

and $N_{all} = \dfrac{\text{Total number of duplicated genes}}{\text{Total number of single genes}}$

and $\begin{cases} R = 1 \Rightarrow \text{proportion of duplicated genes is identical to the average} \\ R > 1 \Rightarrow \text{proportion of duplicated genes is greater than the average} \\ \text{R} < 1 \Rightarrow \text{proportion of duplicated genes is smaller than the average} \end{cases}$

This ratio was computed for each WGD, giving three ratio values for each metabolism pathway: R_{recent} (recent WGD), $\text{R}_{\text{intermediary}}$ (intermediary WGD) and R_{old} (old WGD). Results for the more represented pathways are shown in Table S12.

8 Protein domain analysis

8.1 Detection of protein domains

InterProScan was run against all human, *Arabidopsis thaliana*, *Paramecium tetraurelia*, *Plasmodium falciparum, Tetrahymena thermophila, Drosophila melanogaster, Neurospora crassa, Dictyostelium discoideum, Thalassiosira pseudonana* and *Cyanidioschyzon merolae* proteins as described earlier[29]. Matches, which fulfilled the following criteria, were retained :

- match is tagged as “True Positive” by InterProScan (status=T) ;
- match with an e-value less or equal to 10^{-1}.

2111 InterPro domains (with IPR number) were found in *Paramecium,* and correspond to 18,018 *Paramecium* proteins. So, about 45% of *Paramecium* proteins have referenced InterPro domains.

8.2 Estimate of the proportion of duplicated genes for each domain

At each duplication level, the number of duplicated and of single *Paramecium* proteins was counted for each InterPro domain (IPR). Then the ratio R, which permits comparison of the proportion of duplicated proteins matching a given IPR to the proportion of duplicated proteins in the *Paramecium* genome, were computed for each IPR as follows:

$$R = \frac{R_{IPR}}{R_{WGD}}$$

$$\text{with, } R_{IPR} = \frac{\text{Number of } \textit{Paramecium} \text{ duplicated proteins matching this IPR domain}}{\text{Number of } \textit{Paramecium} \text{ single proteins matching this IPR domain}}$$

$$\text{and } R_{WGD} = \frac{\text{Total number of } \textit{Paramecium} \text{ duplicated proteins}}{\text{Total number of } \textit{Paramecium} \text{ single proteins}}$$

$$\text{and } \begin{cases} R = 1 \Rightarrow \text{proportion of duplicated proteins matching this IPR domain is identical to the average} \\ R > 1 \Rightarrow \text{proportion of duplicated proteins matching this IPR domain is higher than the average} \\ R < 1 \Rightarrow \text{proportion of duplicated proteins matching this IPR domain is lower than the average} \end{cases}$$

This ratio R was computed for each WGD, providing three ratio values: R_{recent} (recent WGD), $R_{intermediary}$ (intermediary WGD) and R_{old} (old WGD).

From the data used to compute the ratio R (as previously), a Chi^2 test (1 degree of freedom) was used to compare the expected distribution of gene numbers against the observed distribution of gene numbers. A p-value was deduced from this test for each domain. If the p-value is less or equal to 5%, we consider the duplicated gene number to be significantly different from the single gene number.
The combination of the ratio R and the p-value permit us to know if a domain is over-represented, under-represented or represented like the average of the domains.

9 Protein complex analysis

To identify complexes in the *Paramecium* proteome, we used two known lists of complexes in *Saccharomyces cerevisiae*[30, 31], and remapped theses complexes on the *Paramecium* proteome using orthologous links.

9.1 Identification of orthologous genes

For each yeast gene, we retained its best reciprocal match (based on scores of Smith-Waterman alignments) in the *Paramecium* proteome (see 5.1). To avoid disturbances of the BRH criteria, we used ancestral proteins (for each WGD event) in order to pair yeast and *Paramecium* peptides. Thus, a given yeast gene could be assigned to one or two *Paramecium* genes for the recent WGD, 1 to 4 for the intermediary WGD and 1 to 8 for the old WGD. In the same way, we used yeast ancestral peptides (i.e. we aggregated duplicated genes found in yeast[32]). We obtained:

- Recent WGD: 943 orthologous links between ancestral proteins, and after splitting yeast duplicated genes, we obtained an orthologous relation for 1,070 yeast peptides.
- Intermediary WGD: 849 pairwise orthologous links for 973 yeast peptides.
- Old WGD: 583 pairwise orthologous links for 675 yeast peptides.

9.2 Identification of complexes for each WGD

To identify protein complexes in the *Paramecium* proteome, we used orthologous relations. Each yeast protein involved in a complex was kept if and only if it had an orthologous protein in *Paramecium*. We focused on complexes of two or more *Paramecium* proteins.

From the MIPS catalogue which contains 1,602 yeast complexes, we remapped on the *Paramecium* proteome:

- 599 protein complexes for the recent WGD
- 562 protein complexes for the intermediary WGD
- 443 protein complexes for the old WGD

We used a second list of protein complexes identified by ref[31] which contains 422 protein complexes with at least 2 peptides in its core. On this catalogue, we remapped:

- 109 protein complexes for the recent WGD
- 101 protein complexes for the intermediary WGD
- 61 protein complexes for the old WGD

9.3 Stoichiometry analysis

We classified inferred *Paramecium* complexes by size and focused on the copy number of each ancestral protein belonging to a complex. We then computed for each size the proportion of protein complexes with perfect stoichiometry in copy number (Tables S9 and S10). To assign significance to the observed stoichiometry, we randomized protein complex composition. For each WGD, we used the initial set of proteins belonging to a complex. Using these sets, we randomly generated as many protein complexes of each size as in the observed data, and performed one million random samples.

9.4 Gene retention analysis

We define the retention level of proteins belonging to a complex as a ratio. This retention ratio was computed as follow:

$$R = \frac{R_{COMP}}{R_{WGD}}$$

$$\text{with, } R_{COMP} = \frac{\text{Number of } \textit{Paramecium} \text{ duplicated proteins belonging to a complex}}{\text{Number of } \textit{Paramecium} \text{ single proteins belonging to a complex}}$$

$$\text{and } R_{WGD} = \frac{\text{Total number of } \textit{Paramecium} \text{ duplicated proteins}}{\text{Total number of } \textit{Paramecium} \text{ single proteins}}$$

$$\text{and} \begin{cases} R = 1 \Rightarrow \text{proportion of duplicated proteins belonging to a complex is identical to the average} \\ R > 1 \Rightarrow \text{proportion of duplicated proteins belonging to a complex is higher than the average} \\ R < 1 \Rightarrow \text{proportion of duplicated proteins belonging to a complex is lower than the average} \end{cases}$$

9.5 Known Interpro interacting domains

We observed a number of Interpro domains that are systematically amplified above the mean across all amplification events (Supplementary File F01). We found that many signalling molecules, such as protein kinases and their substrates, are always amplified, as previously observed for other WGDs. However, not all signalling systems are equally amplified. *Paramecium* seems to have preferentially retained the dual Histidine kinase / response regulators that are frequently involved in signal transduction in bacteria and protozoans. Importantly, we observed that the two components of the signalling system covary in their retention rate across the duplication events (Fig S13). This co-variation is seen with other well-characterized systems, like small GTPases and their associated GTPase Activating Proteins (Fig S13).

9.6 Other complexes

Among the few ciliate-specific complexes that were previously known, we observed the same over-retention pattern. The Trichocyst Matrix Proteins (TMPs) encoded by a large multi-gene family, are the major components of the dense crystalline core of voluminous secretory vesicles known as trichocysts. Gene silencing experiments showed that perturbation of the stoichiometry of different TMP sub-families compromises their assembly, leading to trichocysts that cannot be secreted. We find that the TMP family arose essentially through the WGDs and these genes are over-retained in all duplications (Supplementary Table S13). We also identified a large number of proteins constitutive of cilia, through orthology with other organisms. Most of these proteins are engaged in multi-protein complexes and all work together to assure the biogenesis and motility of this important organelle. The set of ciliary genes is over-retained in the recent duplication (Supplementary Fig. S15).

10 Detection of Paramecium proteins involved in ciliary and basal body function

Two sets of data were used to detect Paramecium putative proteins involved in ciliary and basal body function: flagellar proteins from the algae *Chlamydomonas reinhardtii*[33] identified by a proteomic approach and ciliary and basal body proteins identified by a comparative genomic approach[34]. The proteins in these sets were compared by BLAST and the results filtered, after careful validation of the alignments using minimal thresholds of 27% for the identity and of 50% for the coverage. The hits sample was extended using a score calculated by fixing the product of these two numbers (1350) as a threshold, then by adding hits for which the score was between 1100 and 1350, but with a minimum of 46% coverage. The two sets contain a majority of proteins detected in *Paramecium* (Fig S14). A clear conservation of major proteins is found in all three species.

11 Searching for asymmetry in the rate of evolution of gene duplicates

We analyzed the rate of protein evolution for pairs of duplicates from the recent and intermediary WGDs for which an outgroup from a more ancient WGD was available. We excluded all pseudogenes detected previously with genewise[11] (see above). For each triplet (one pair of duplicates and an outgroup) we computed the protein alignment with ClustalW. To exclude partial genes that might correspond to recent pseudogenes, we only retained gene pairs for which the protein alignment covers 90% of the length of the longest sequence. The final dataset contains 2297 pairs of recent duplicates with an outgroup from the intermediary WGD, and 293 pairs of intermediary duplicates with an outgroup from the old WGD.
A likelihood-ratio test was performed to determine whether the two duplicates evolved at the same rate or not. The likelihood of the phylogenetic tree was estimated with PAML[26], under two models of sequence evolution: in the first model we assumed that the two duplicates evolved at the same rate (molecular clock). In the second model we assumed that the two duplicates evolved at different rates (no molecular clock). Twice the difference in log-likelihood between the two models follows a χ^2 distribution (one degree of freedom). We detected 366 (15.9%) pairs of recent duplicates and 62 (21.2%) pairs of intermediary duplicates for which the "molecular clock" model was rejected with a p-value lower than 5% (asymmetric pairs). With this p-value threshold, one expects by chance to detect 5% of false positive. Hence we conclude that 10.9% of recent duplicates and 16.2% of intermediary duplicates show an

asymmetric rate of evolution. The difference in the frequency of asymmetric pairs between these WGD events is significant ($\chi^2 = 7.55$ $p<0.01$).

Recent duplicates are more recent than intermediary duplicates, and hence are less divergent. The power of the likelihood-ratio test depends on the evolutionary distance between sequences (if sequences are very closely related, it may not be possible to detect a significant difference in their rate of evolution). To test whether the difference between the recent and the intermediary WGD in the frequency of asymmetric pairs is due to a lack of power of the likelihood-ratio test, we split the data set according to the rate of evolution of the slowly evolving copy (hereafter named 'copy1'; the fast-evolving copy being name 'copy2'). As shown in Table S14, the frequency of asymmetric pairs is always higher for intermediary duplicates than for recent duplicates, whatever their rate of evolution. Hence, the difference in the frequency of asymmetric pairs between recent and intermediary duplicates is not due to a lack of power of the likelihood-ratio test.

We computed the rate of evolution of copy1 and copy2, for asymmetric or non-asymmetric pairs of duplicates. For copy1, there is no significant difference between the substitution rates of asymmetric or non-asymmetric pairs, whereas for copy2, the substitution rate is two times higher for asymmetric than for non-asymmetric pairs (Table S15). We therefore conclude that the asymmetries detected are essentially due to an increase in the rate of evolution in copy2, and not to a slow-down in the rate of evolution of copy1.

12 Estimation of the number of pseudogenes

To detect recent pseudogenes annotated as genes, each duplicated gene was aligned against the genomic loci of its paralog using genewise[11]. Alignment indels (insertion or deletion event) were extracted from the 24,052 genewise alignments. These 6,784 indels were shared by 2,627 genes. After removing indels due to consensus sequence potential errors, we obtained a list of 1,499 candidate pseudogenes, about 6% of the total amount of duplicated genes.

13 Dating of the recent duplication

To date the recent WGD, we analyzed all gene families containing at least two *P. tetraurelia* paralogs resulting from the recent duplication, and for which homologs in other species of the *Paramecium* genus were available (N= 15 gene families). In phylogenetic trees, *P. tetraurelia* paralogs resulting from the recent WGD are indicated with the suffix A and B. Paralogs resulting from more ancient duplications are numbered (e.g. *P. tetraurelia* 1A, 1B, 2A, 2B, etc.). For other species, GenBank accession numbers are indicated in the figure legend, except for *Tetrahymena* proteins that were downloaded from The Institute for Genomic Research (http://www.tigr.org/tdb/e2k1/ttg/).

Eight of these families include a sequence from *P. caudatum* or *P. multimicronucleatum*. In all cases, phylogenetic trees indicate with strong bootstrap support that the recent WGD occurred after the divergence of *P. tetraurelia* from *P. caudatum* and *P. multimicronucleatum* (Fig. S23.2, S23.6, S23.10, S23.11, S23.12, S23.13, S23.14, S23.15).

P. tetraulia belongs to a complex of 15 sibling species (the *Paramecium aurelia* complex[35,36]) : *P. primaurelia, P. biaurelia, P. triaurelia, P. tetraurelia, P. pentaurelia, P. sexaurelia, P. septaurelia, P. octaurelia, P. novaurelia, P. decaurelia, P. undecaurelia, P. dodecaurelia, P. tredecaurelia, P. quadecaurelia, P. sonneborni*. Nine families with at least one of these species have been analyzed.

The Rab7a and Rab7b paralogous genes result from the recent WGD. We identified the two genes, both in and *P. tetraurelia* and *P. octaurelia*, which demonstrates that the recent WGD predates the divergence of *P. octaurelia* and *P. tetraurelia* (Fig. S23.1). This conclusion is also supported by the four other gene families analyzed in *P. octaurelia* (Fig. S23.2a, S23.4, S23.5, S23.7).
Two paralogous cytosolic HSP70 genes resulting from the recent WGD are present in *P. tetraurelia.* Cytosolic HSP70 genes have been sequenced in several strains of each species of the *P. aurelia* complex[36]. This data set includes both some paralogous and some allelic sequences. The phylogenetic tree indicates with strong bootstrap support that the recent WGD occurred before the divergence of *P. tetraurelia* from *P. triaurelia, P. septaurelia, P. octaurelia,* and *P. dodecaurelia*. The rest of the tree is poorly resolved, but suggests that the recent WGD occurred before the divergence of all the species of the aurelia complex (Fig S23.2a). Interestingly, in *P. sexaurelia* and *P. biaurelia*, we identified two sets of sequences that cluster with each of the two *P. tetraurelia* paralogs (Fig. S23.2b). This demonstrates that the recent WGD predates the divergence of *P. tetraurelia* from *P. sexaurelia* and *P. biaurelia*.
The three other families for which *P. primaurelia* sequences are available also indicate that the recent WGD occurred before the divergence of *P. tetraurelia* from *P. primaurelia* (Fig. S23.3, S23.8, S23.9).
The additional data available for *P. triaurelia*, and *P. jenningsi* also suggests that the recent WGD occurred before their divergence from of *P. tetraurelia* (Fig. S23.6).
All these lines of evidence indicate that the radiation of the *P. aurelia* complex occurred shortly after the recent WGD.

14 Dating the old and intermediary genome duplications

To date the intermediary and old duplications, we randomly sampled 91 gene families having retained paralogs from both of these WGDs. We then selected families that were suitable for phylogenetic analyses, i.e. for which homologs with reliable alignments were found in *Tetrahymena thermophila* and several other eukaryotes used as outgroups. Among the 27 families that could be analyzed, we found 19 cases where the old duplication appears to be specific to the *Paramecium* lineage and 8 cases (30%) where the old duplication appears to predate the divergence between *T. thermophila* and *P. tetraurelia*. It should be stressed that because of gene conversion events, phylogenetic trees of paralogous genes can only give a lower bound of the duplication date. We therefore propose that the old duplication occurred shortly before the divergence of *Paramecium* and *Tetrahymena*. In agreement with that conclusion, the distributions of the percentage identity between *Paramecium* paralogs and *Paramecium-Tetrahymena* orthologs overlap almost perfectly (Fig. 3b). Nonetheless, an analysis of the conservation of synteny between *Paramecium* and *Tetrahymena* will be necessary to formally demonstrate this dating. For all the 27 families analyzed from the intermediary duplication, the event appears to be specific to the *Paramecium* lineage.

References

1. Keller, A. M. & Cohen, J. An indexed genomic library for Paramecium complementation cloning. J Eukaryot Microbiol 47, 1-6 (2000).
2. Skouri, F. & Cohen, J. Genetic approach to regulated exocytosis using functional complementation in Paramecium: identification of the ND7 gene required for membrane fusion. Mol Biol Cell 8, 1063-71 (1997).
3. Tamm, S. L., Sonneborn, T. M. & Dippell, R. V. The role of cortical orientation in the control of the direction of ciliary beat in Paramecium. J Cell Biol 64, 98-112 (1975).
4. Preer, L. B., Hamilton, G. & Preer, J. R., Jr. Micronuclear DNA from Paramecium tetraurelia: serotype 51 A gene has internally eliminated sequences. J Protozool 39, 678-82 (1992).
5. Batzoglou, S. et al. ARACHNE: a whole-genome shotgun assembler. Genome Res 12, 177-89 (2002).
6. Smit, A., Hubley, R & Green, P. *RepeatMasker Open-3.0* 1996-2004 http://www.repeatmasker.org.
7. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573-80 (1999).
8. Roest Crollius, H. et al. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. Nat Genet 25, 235-8 (2000).
9. Jaillon, O. et al. Genome-wide analyses based on comparative genomics. Cold Spring Harb Symp Quant Biol 68, 275-82 (2003).
10. Bairoch, A. et al. The Universal Protein Resource (UniProt). Nucleic Acids Res 33, D154-9 (2005).
11. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. Genome Res 14, 988-95 (2004).
12. Parra, G., Blanco, E. & Guigo, R. GeneID in Drosophila. Genome Res 10, 511-5 (2000).
13. Korf, I. Gene finding in novel genomes. BMC Bioinformatics 5, 59 (2004).
14. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20, 2878-9 (2004).
15. Parra, G. et al. Comparative gene prediction in human and mouse. Genome Res 13, 108-17 (2003).
16. Porcel, B. M. et al. Numerous novel annotations of the human genome sequence supported by a 5'-end-enriched cDNA collection. Genome Res 14, 463-71 (2004).
17. Castelli, V. et al. Whole genome sequence comparisons and "full-length" cDNA sequences: a combined approach to evaluate and improve Arabidopsis genome annotation. Genome Res 14, 406-13 (2004).
18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J Mol Biol 215, 403-10 (1990).
19. Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci 13, 477-8 (1997).
20. Kent, W. J. BLAT--the BLAST-like alignment tool. Genome Res 12, 656-64 (2002).
21. NCBI. http://www.ncbi.nlm.nih.gov/.
22. Lee, Y. et al. The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Res 33, D71-4 (2005).
23. Howe, K. L., Chothia, T. & Durbin, R. GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. Genome Res 12, 1418-27 (2002).
24. Zagulski, M. et al. High coding density on the largest Paramecium tetraurelia somatic chromosome. Curr Biol 14, 1397-404 (2004).
25. Finn, R. D. et al. Pfam: clans, web tools and services. Nucleic Acids Res 34, D247-51 (2006).
26. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13, 555-6 (1997).
27. Claudel-Renard, C., Chevalet, C., Faraut, T. & Kahn, D. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res 31, 6633-9 (2003).

28. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. Nucleic Acids Res 30, 42-6 (2002).
29. Zdobnov, E. M. & Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847-8 (2001).
30. MIPS. http://mips.gsf.de/genre/proj/yeast/searchCatalogFirstAction.do?style=catalog.xslt&table=CELLULAR_COMPLEXES&db=CYGD.
31. Gavin, A. C. et al. Proteome survey reveals modularity of the yeast cell machinery. Nature (2006).
32. Byrne, K. P. & Wolfe, K. H. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res 15, 1456-61 (2005).
33. Pazour, G. J., Agrin, N., Leszyk, J. & Witman, G. B. Proteomic analysis of a eukaryotic cilium. J Cell Biol 170, 103-13 (2005).
34. Li, J. B. et al. Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. Cell 117, 541-52 (2004).
35. Coleman, A. W. Paramecium aurelia revisited. J Eukaryot Microbiol 52, 68-77 (2005).
36. Hori, M., Tomikawa, I., Przybos, E. & Fujishima, M. Comparison of the evolutionary distances among syngens and sibling species of Paramecium. Mol Phylogenet Evol 38, 697-704 (2006).
37. Baroin, A., Prat, A. & Caron, F. Telomeric site position heterogeneity in macronuclear DNA of Paramecium primaurelia. Nucleic Acids Res 15, 1717-28 (1987).
38. Stajich, J. E. et al. The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12, 1611-8 (2002).
39. Coulondre, C., Miller, J. H., Farabaugh, P. J. & Gilbert, W. Molecular basis of base substitution hotspots in Escherichia coli. Nature 274, 775-80 (1978).
40. Ponger, L. & Li, W. H. Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. Mol Biol Evol 22, 1119-28 (2005).

Supplementary Figures

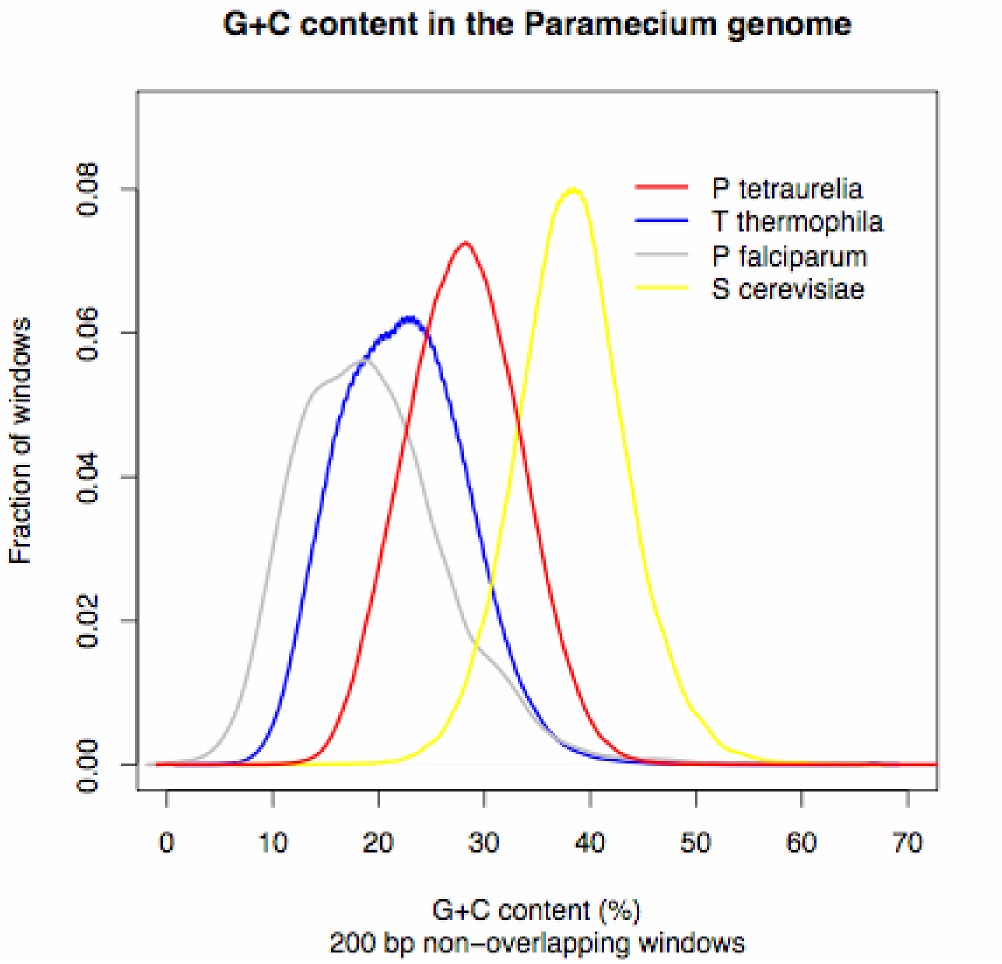


Figure S1. Histogram of GC content. G+C content was calculated using 200 bp non-overlapping windows for the *P. tetraurelia* macronuclear draft genome as well as the *Plasmodium falciparum* genome (http://plasmodb.org), the *Saccharomyces cerevisiae* genome (http://www.yeastgenome.org/) and the *Tetrahymena thermophila* macronuclear draft genome (http://www.tigr.org). The symmetric shape of the *Paramecium* curve is consistent with the absence of any highly repeated sequences.

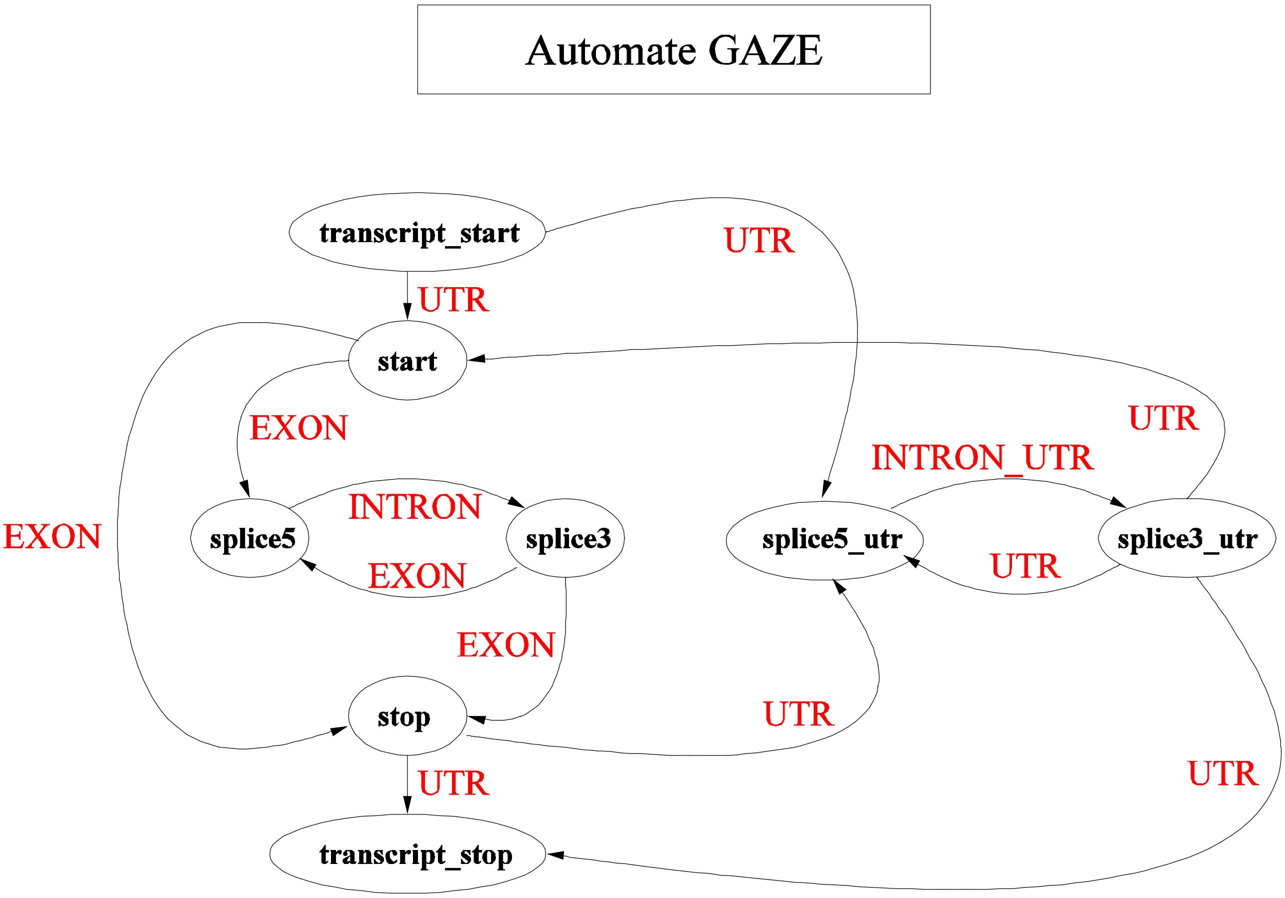


Figure S2. Simplified representation of the GAZE automaton designed to build *Paramecium* gene models.

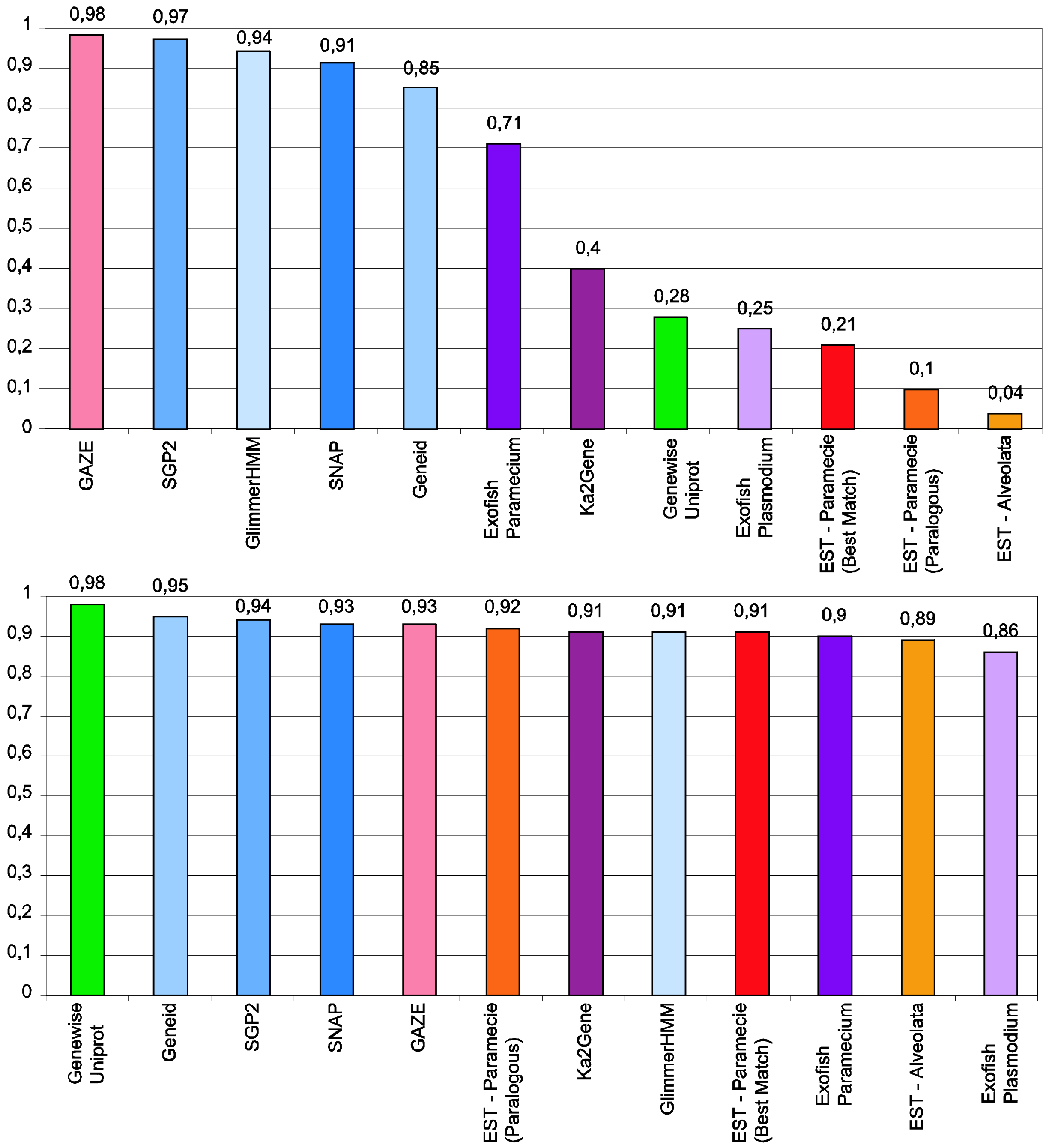
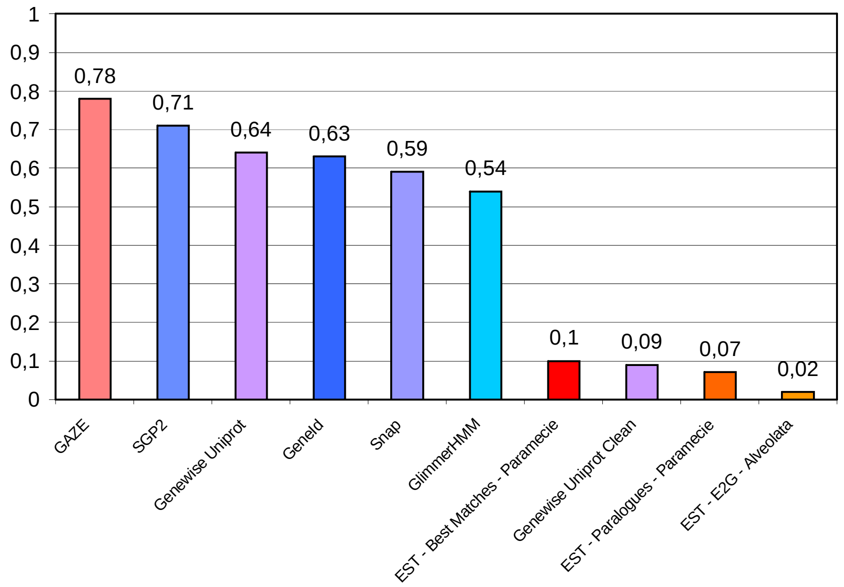


Figure S3. A. Sensitivity accuracy at the nucleotide level of each prediction method. B. Specificity accuracy at the nucleotide level of each prediction method.



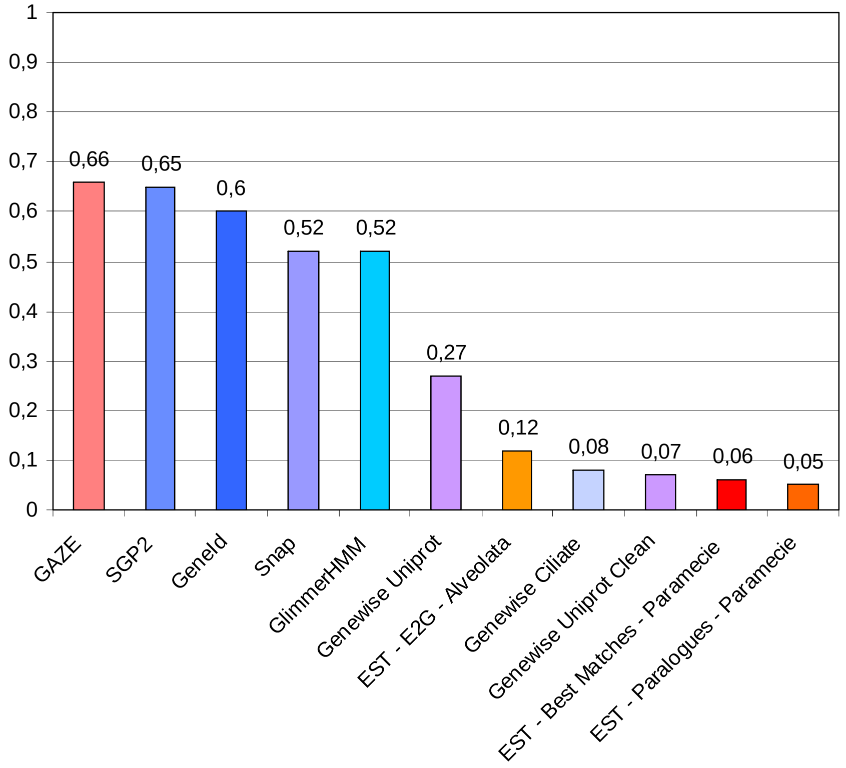


Figure S4. Only methods with exon boundary predictions are shown here. A. Sensitivity accuracy at the exon level of each prediction method. B. Specificity accuracy at the exon level of each prediction method.

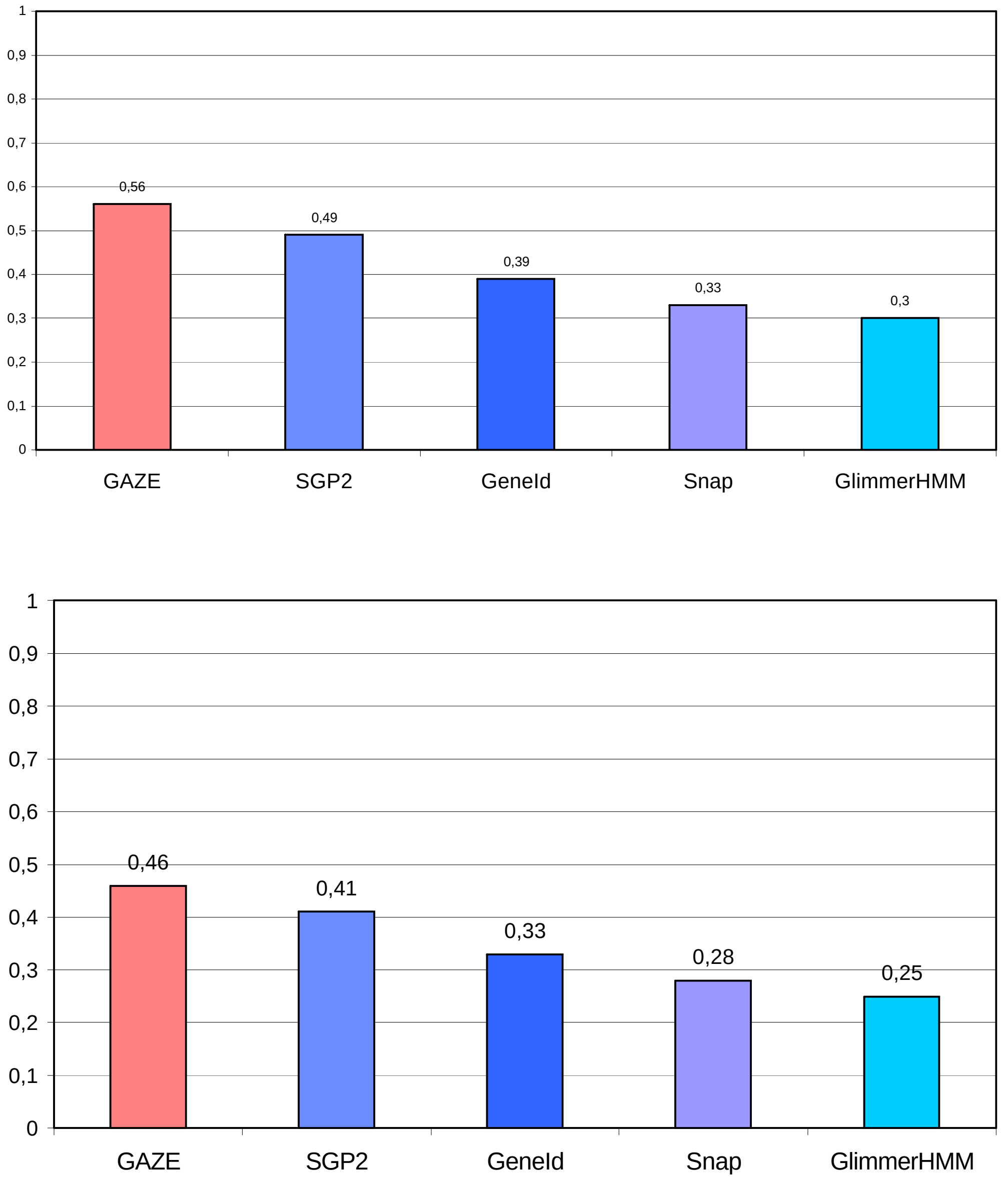
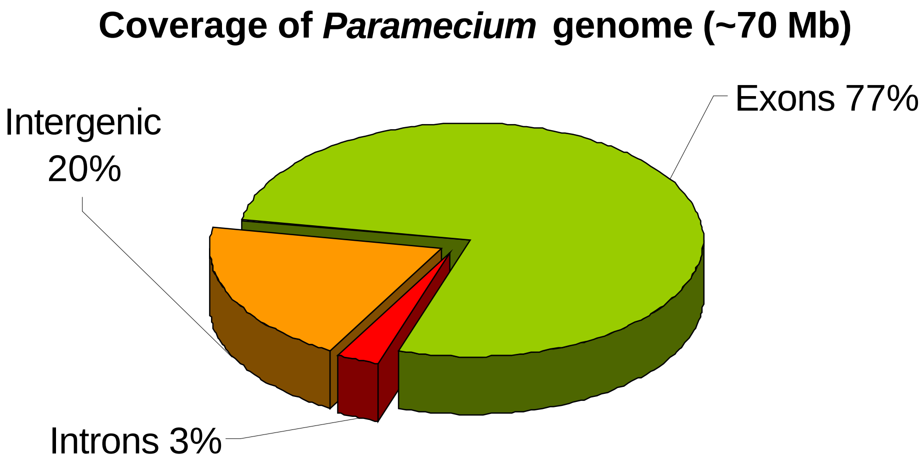


Figure S5. Only methods that provide complete gene predictions are shown here. A. Sensitivity accuracy at the gene level of each prediction method. B. Specificity accuracy at the gene level of each prediction method.



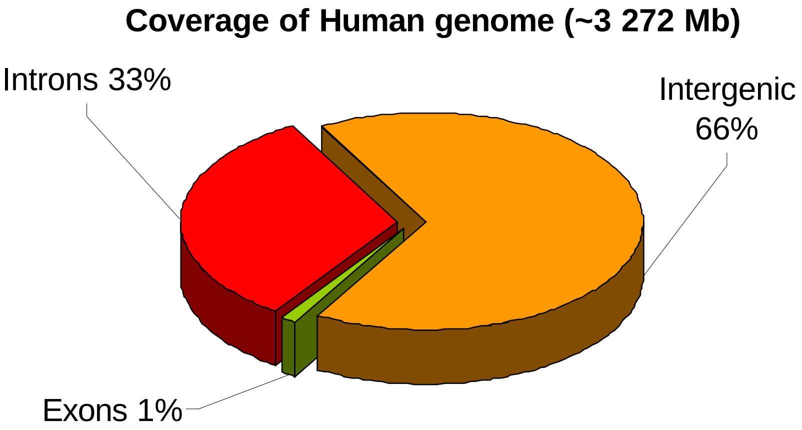


Figure S6. Coverage of coding and non-coding regions in the *Paramecium* genome compared to coverage in the Human genome.

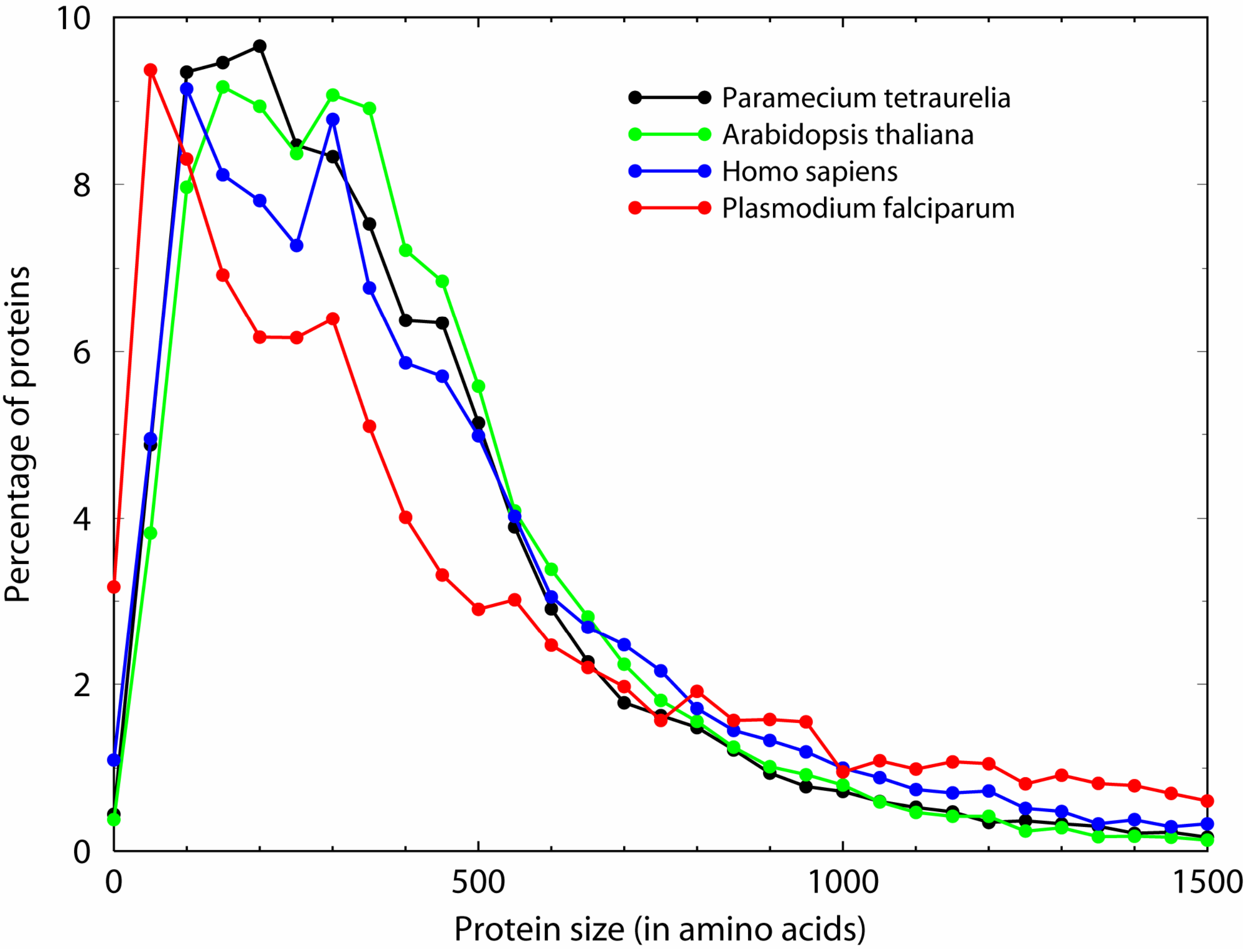


Figure S7. Distribution of *Paramecium* peptide lengths (black) compared to Human (blue), Arabidopsis (green) and Plasmodium (red).

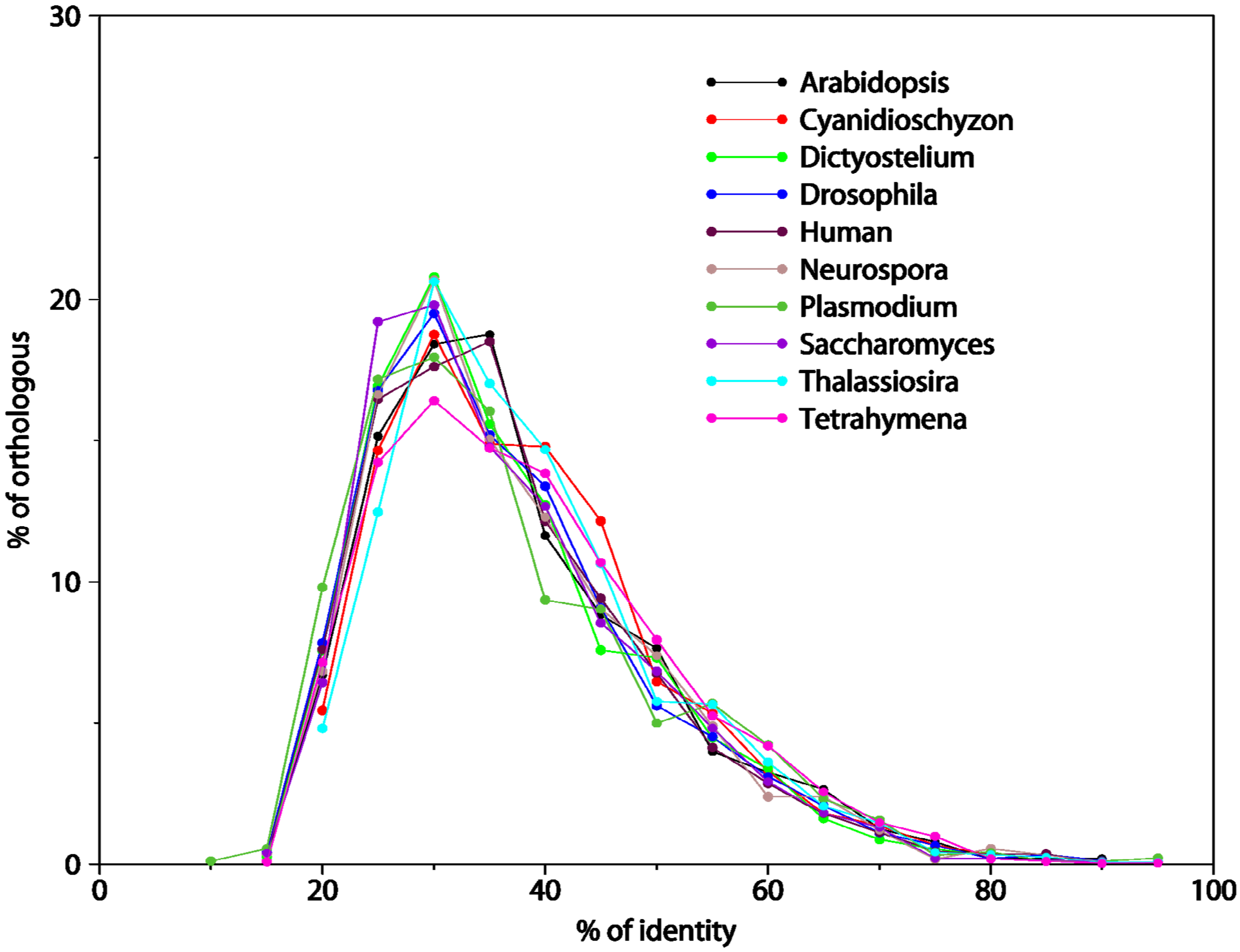
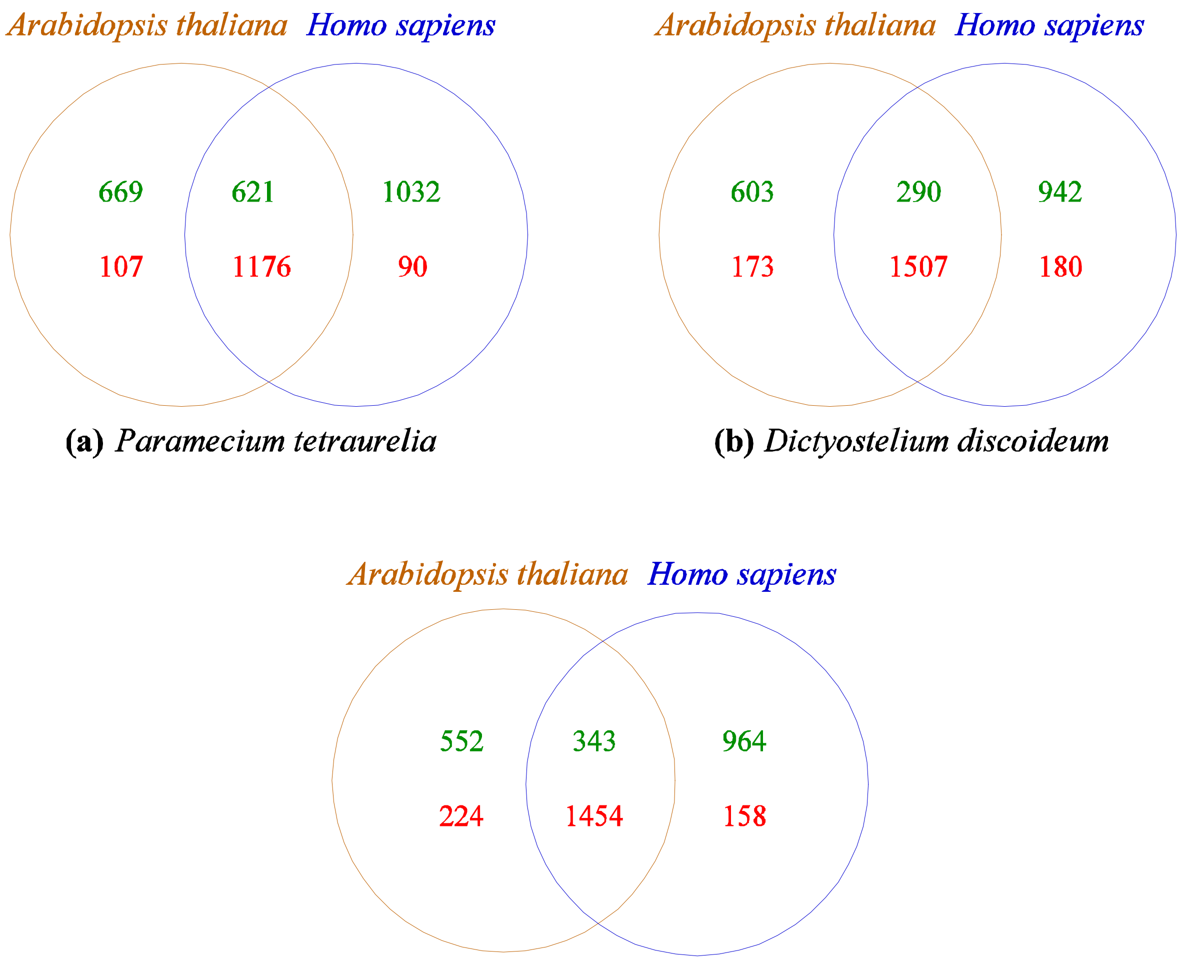


Figure S8. Distribution of the percent identity between pairs of orthologous protein sets.



(a) *Paramecium tetraurelia*

(b) *Dictyostelium discoideum*

(c) *Neurospora crassa*

Figure S9. Distribution of *Paramecium*, *Dictyostelium discoideum* and *Neurospora crassa* Pfam domains (compared species) among *Homo sapiens* and *Arabidopsis thaliana*. The number of Pfam domains present in compared species is in red. The number of Pfam domains absent in compared species is in green. In each group (*A. thaliana* specific, *H. sapiens* specific, and *H.sapiens* - *A. thaliana* common) the sum of red and green numbers is equal to the number of Pfam domains present in this group.

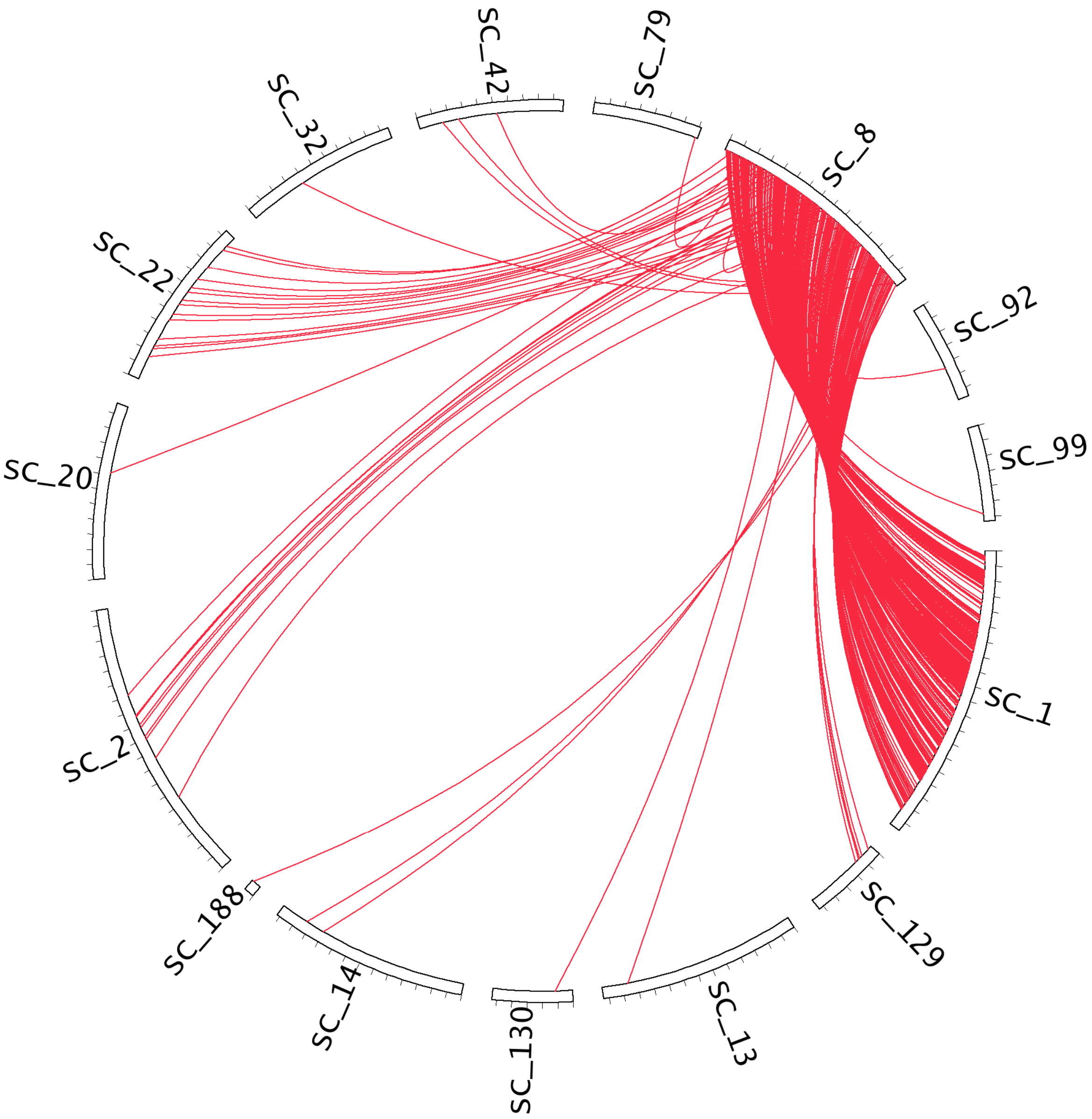


Figure S10. Best Reciprocal Hits (BRH) of scaffold_8 genes. Scaffold_8 and scaffold_1 are paralogous in the recent duplication and share many BRHs, however scaffold_8 shares some BRH links with other scaffolds.

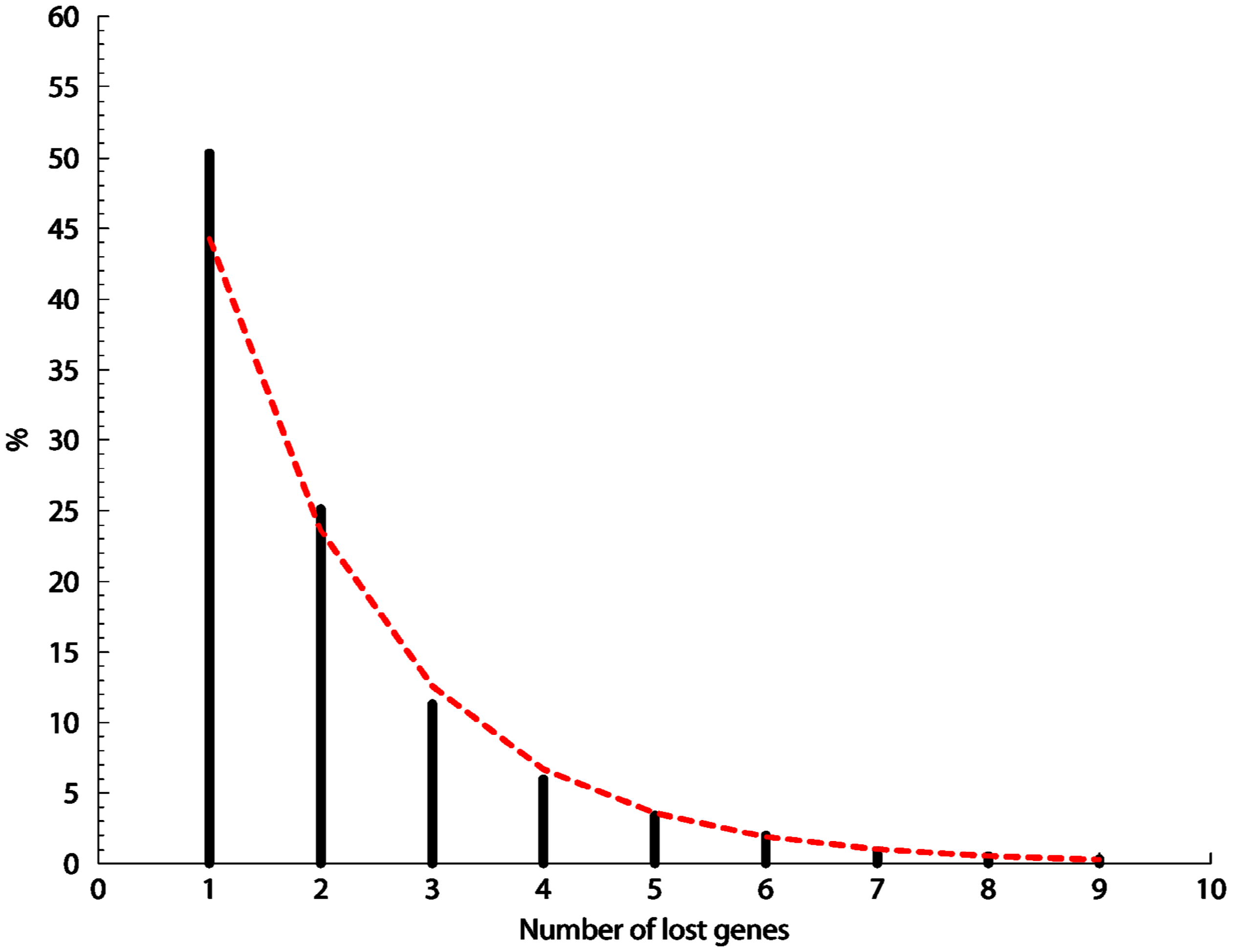


Figure S11. Distribution of the number of deleted genes in recent WGD paralogons. Dotted line: exponential regression CC=0.998.

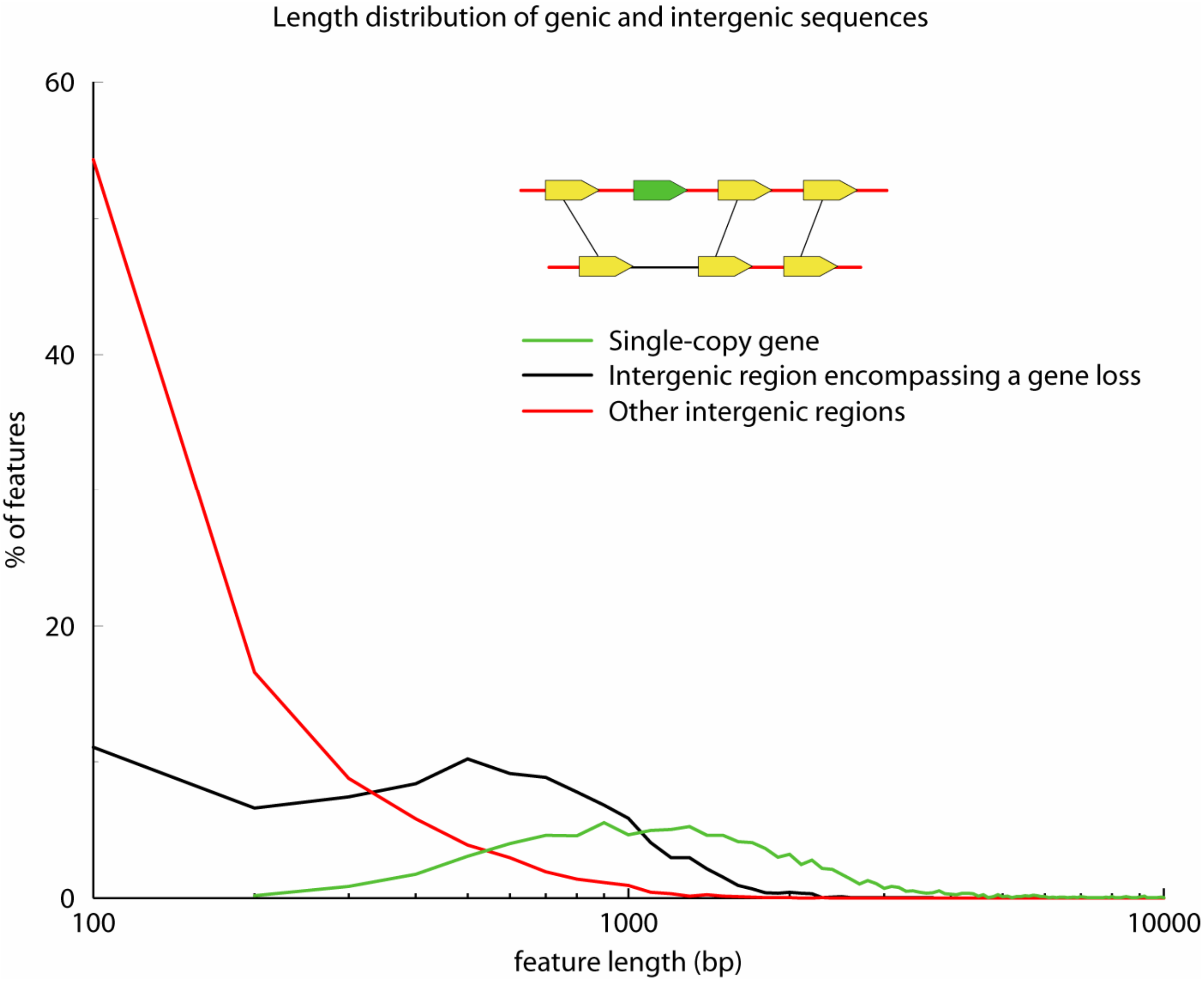


Figure S12. Pseudogenes in intergenic regions. We selected genomic regions situated between two conserved genes after the recent WGD that correspond to one gene in one paralogous chromosome (green curve) and no gene in the other paralogous chromosome (black curve). The diagram shows a plot of the length of these regions versus their percentage, compared with the total number of intergenic regions (red curve). The black curve does not overlap the two others, showing a progressive reduction of size of pseudogenic regions that still contain more sequence remnants than the mean intergenic regions.

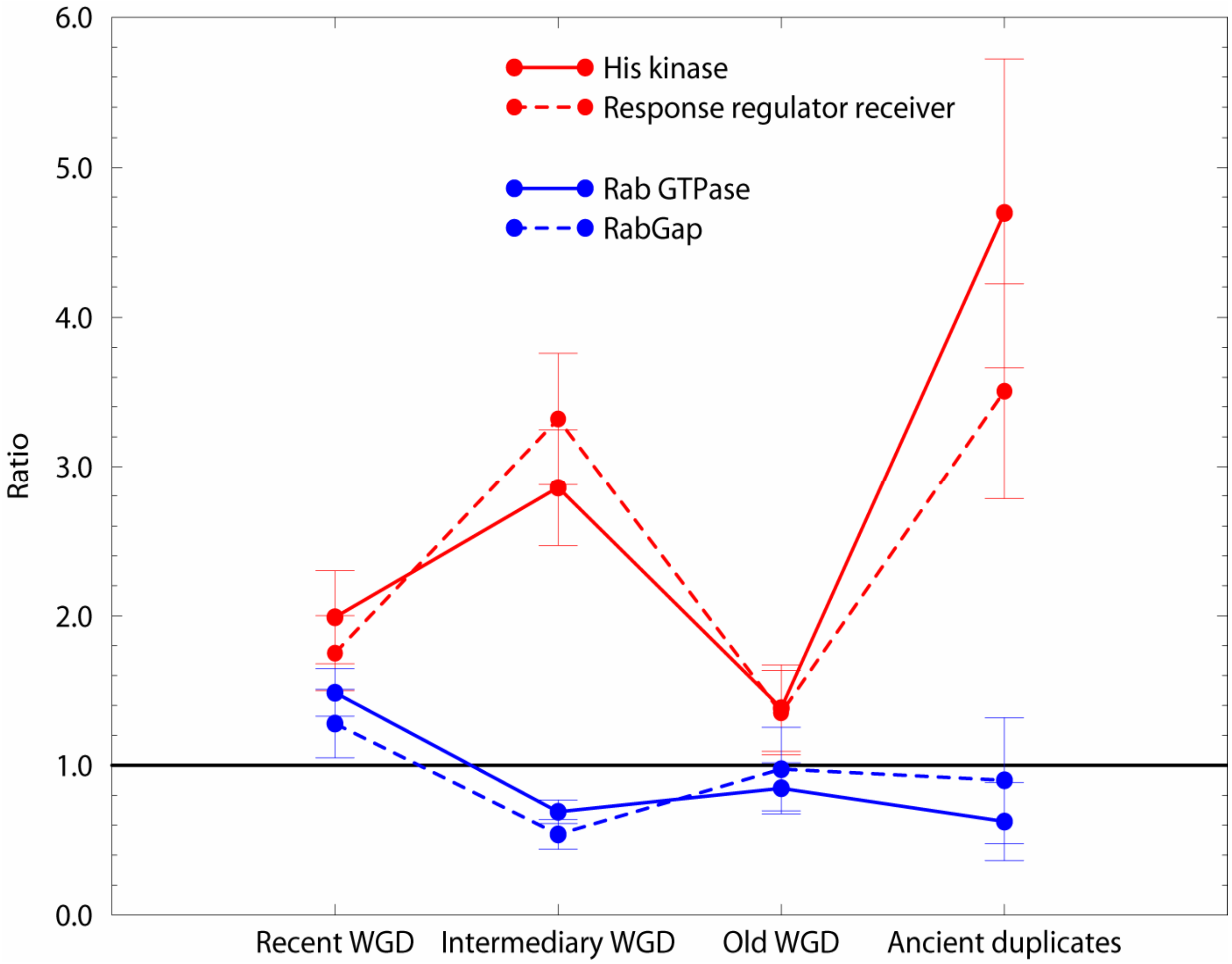


Figure S13. Co-retention of classes of interacting proteins. Red curves : Co-variation of Two-component system constituent numbers across the successive duplications. The numbers of genes coding for Histidine Kinases (solid line) and Response Regulator Receiver (broken line) were estimated by counting the genes containing at least one Interpro46 domain IPR009082 and IPR001789, respectively. The ordinate represents the ratio of observed retained versus non-retained genes at each duplication versus the same ratio for all genes. A value of 1 means that this gene category is retained at the same level as the mean for the whole proteome. Blue curves : Co-variation of Ras GTPases (Rab type) and GTPase Activating Protein (RabGAP type) numbers across the successive duplications. The numbers of genes coding for Rab GTPases (solid line) and RabGAPs (broken line) were estimated by counting the genes containing at least one Interpro domain IPR003579 and IPR000195, respectively. Error bars were computed by bootstrap (1000 replicates).

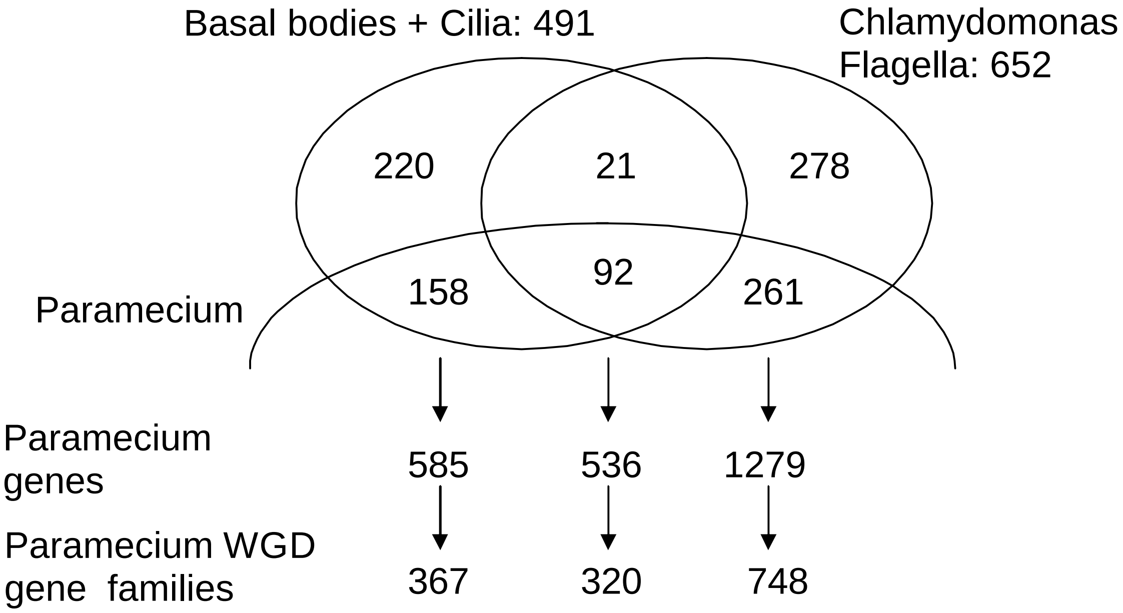

Basal bodies + Cilia: 491
Chlamydomonas Flagella: 652
220
21
278
92
158
261
Paramecium
Paramecium genes
585
536
1279
Paramecium WGD gene families
367
320
748

Figure S14. Comparison of ciliary and flagellar proteomes to Paramecium gene models. Best reciprocal match analysis was used to estimate the presence or absence of each protein of a set in the other sets. The diagram shows the numbers of ciliary and flagellar proteins shared or not in the 3 different species, Homo, Chlamydomonas and Paramecium. Below are indicated the number of genes encoding these proteins, and the number of families generated by WGD to which these genes belong.

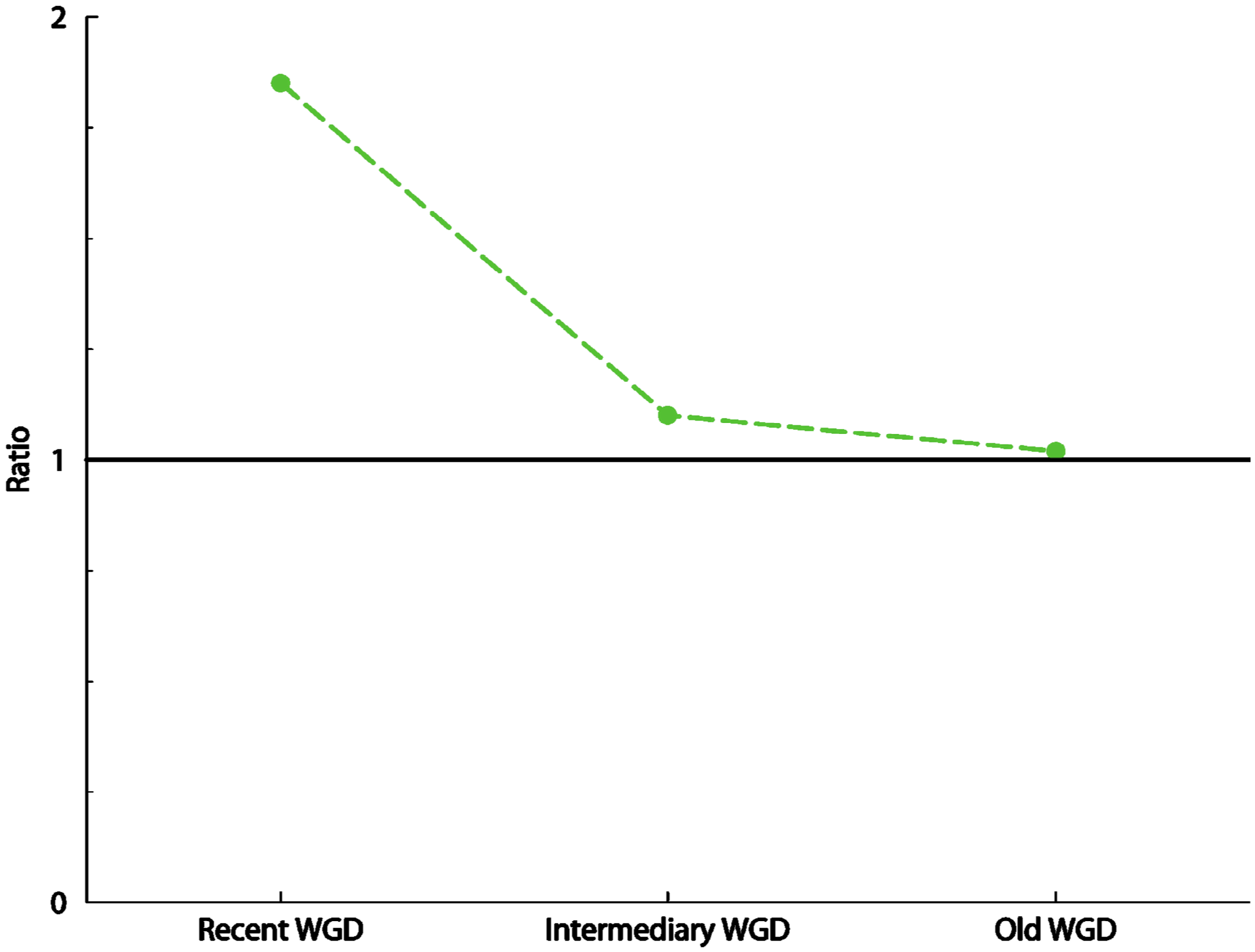


Figure S15. Retention rates for Paramecium genes encoding ciliary and flagellar proteins.

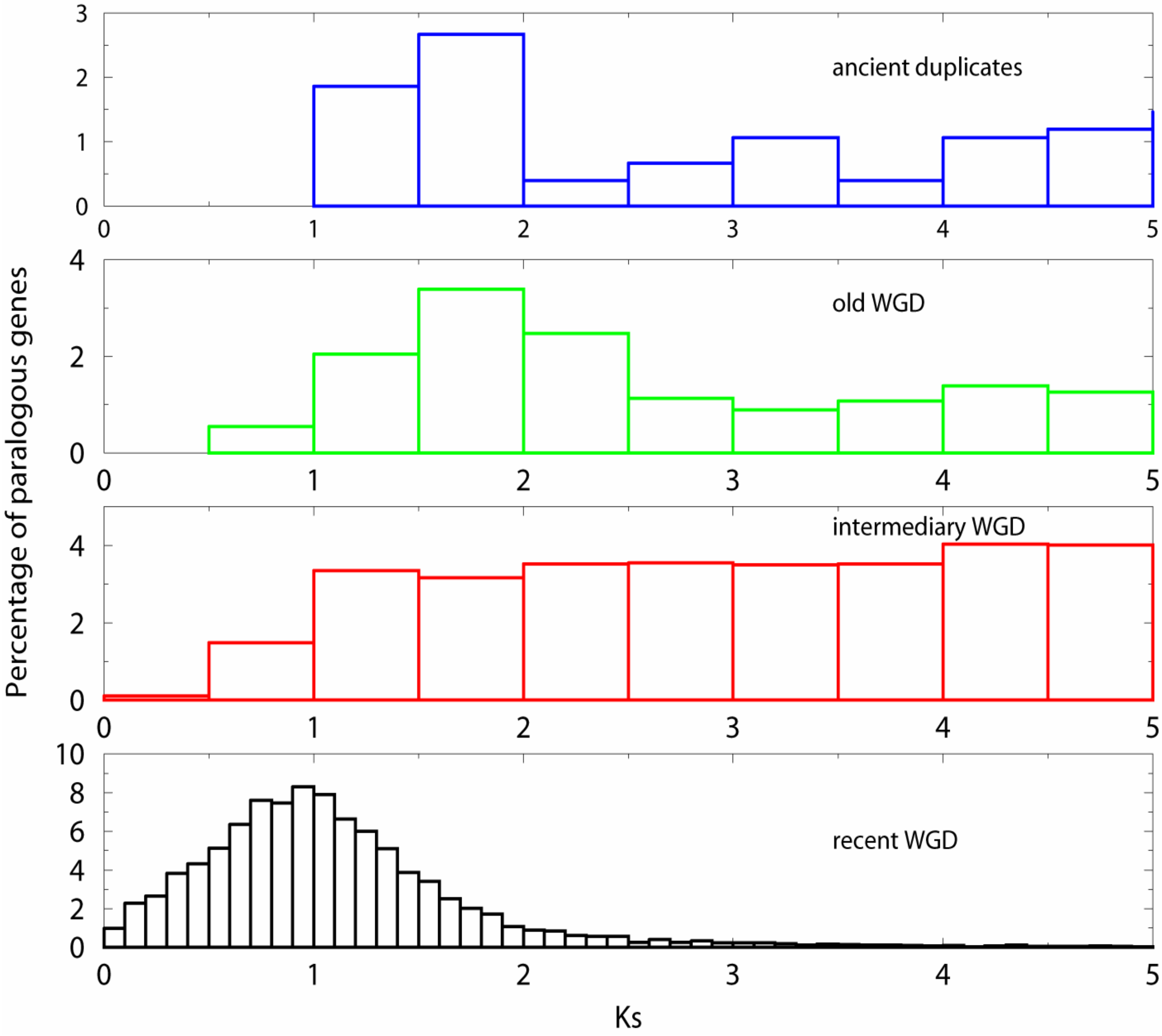


Figure S16. Distribution of Ks values between paralogous genes for each WGD event.

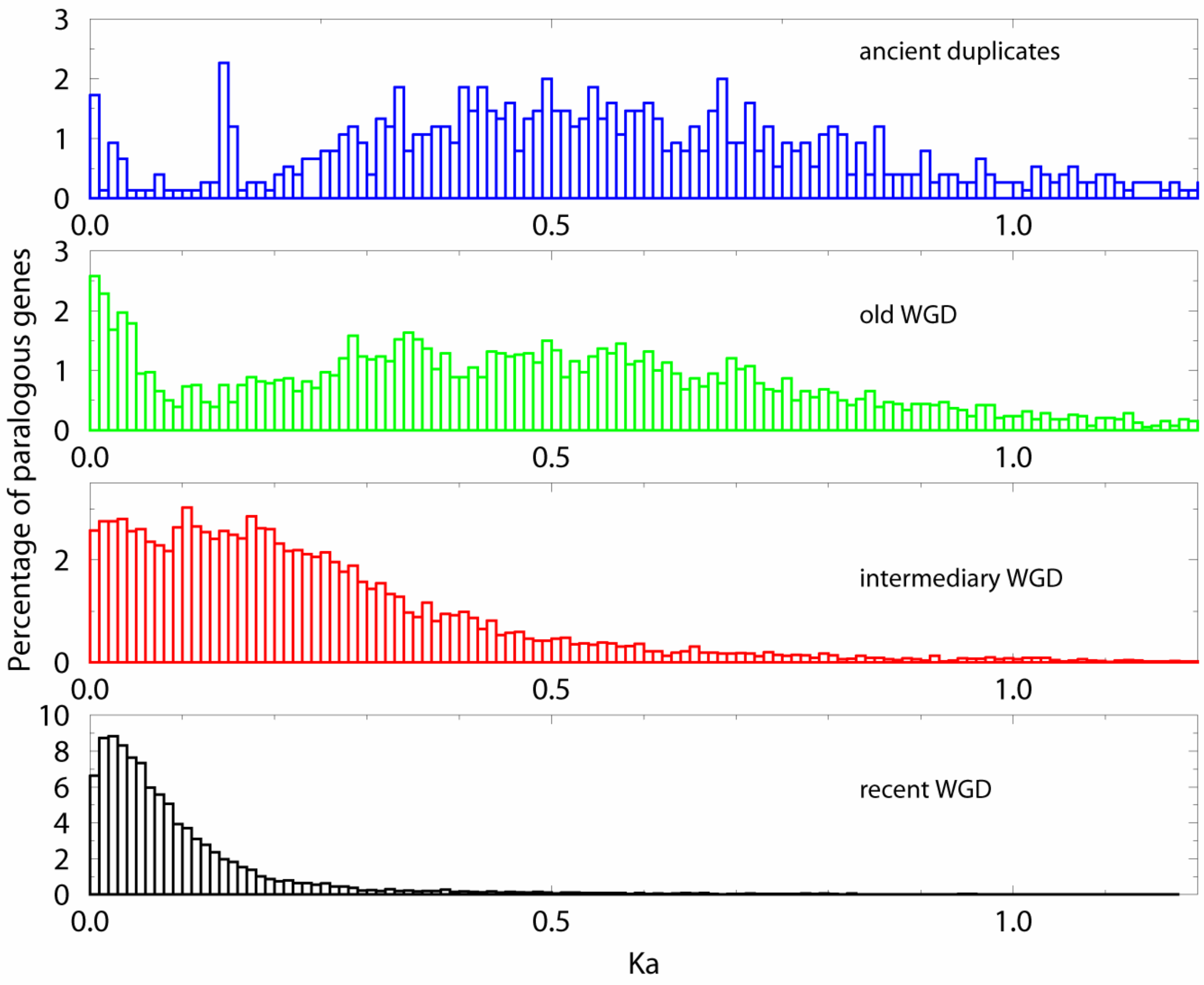


Figure S17. Distribution of Ka values between paralogous genes for each WGD event.

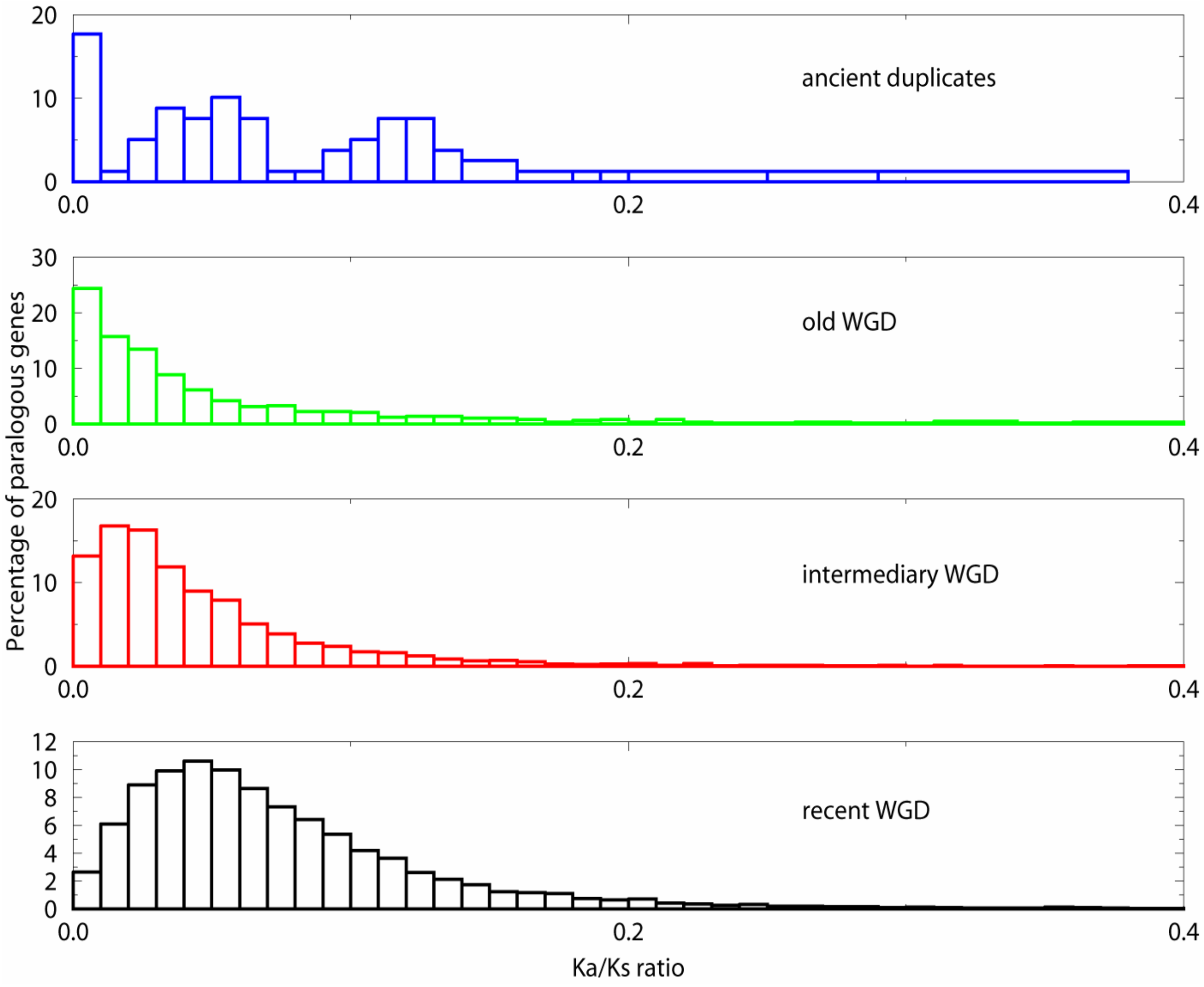
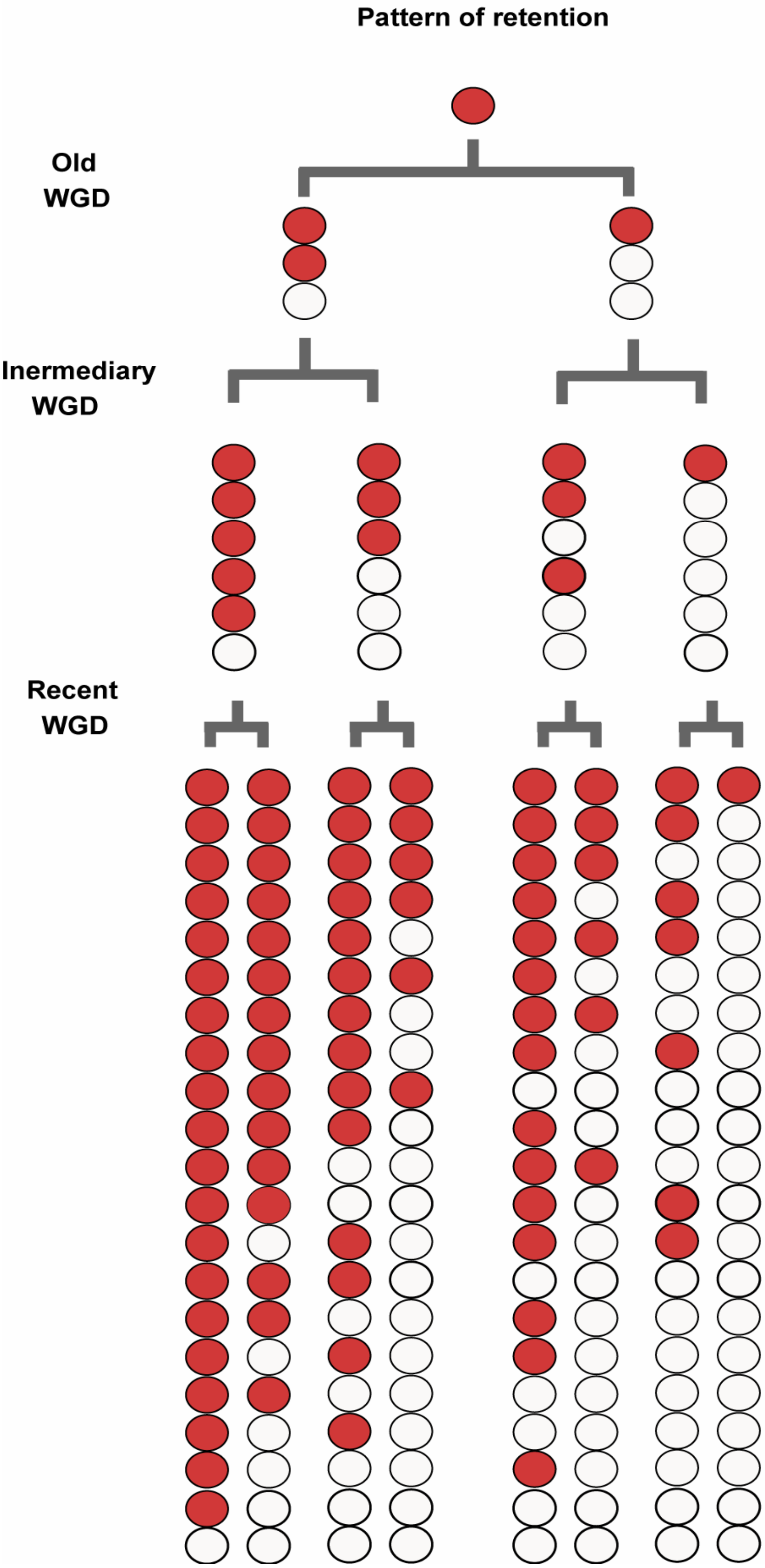


Figure S18. Distribution of Ka/Ks ratio between paralogous genes for each WGD event. Only paralagous genes with a Ks < 5 were considered.



Number of cases	%	Number of genes	%
9234	100	9234	100
765	8.3	1530	15.3
8469	91.7	8469	84.7
NA	NA	NA	NA
107	1.2	428	3.5
343	3.7	1029	8.3
1830	19.8	3660	29.5
315	3.4	630	5.1
6639	71.9	6639	53.6
NA	NA	NA	NA
26	0.3	208	1.1
28	0.3	196	1.0
82	0.9	492	2.6
12	0.1	72	0.4
25	0.3	150	0.8
41	0.4	250	1.1
115	1.2	575	3.1
12	0.13	60	0.3
595	6.4	2380	12.7
52	0.6	208	1.1
117	1.3	468	2.5
33	0.4	132	0.7
4	0.0	16	0.1
865	9.4	2595	13.8
143	1.5	429	2.3
20	0.2	60	0.3
3057	33.1	6114	32.5
370	4.0	740	3.9
55	0.6	110	0.6
3582	38.8	3582	19.1
NA	NA	NA	NA

Figure S19. Successive patterns of deletion. Filled circles: retained genes; white circles: lost genes. After 3 successive WGD, 21 different patterns of retention descend from a unique ancestral gene.

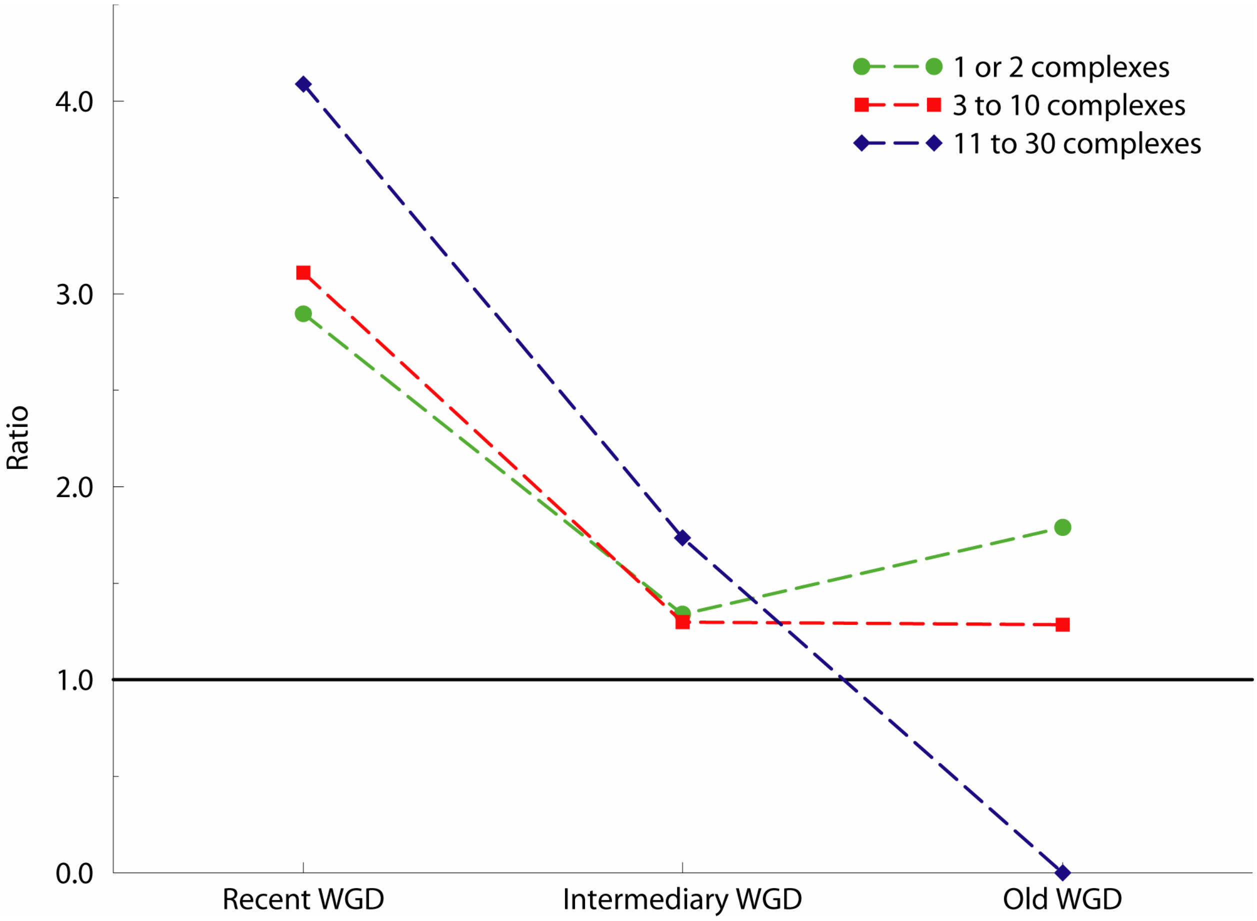


Figure S20. Retention of *Paramecium* proteins belonging to MIPS complexes throughout WGD events, depending on the number of complexes they belong to. Proteins which belong to 1 or 2 complexes are in green, to 3 to 10 complexes in red and to 11 to 30 complexes in blue.

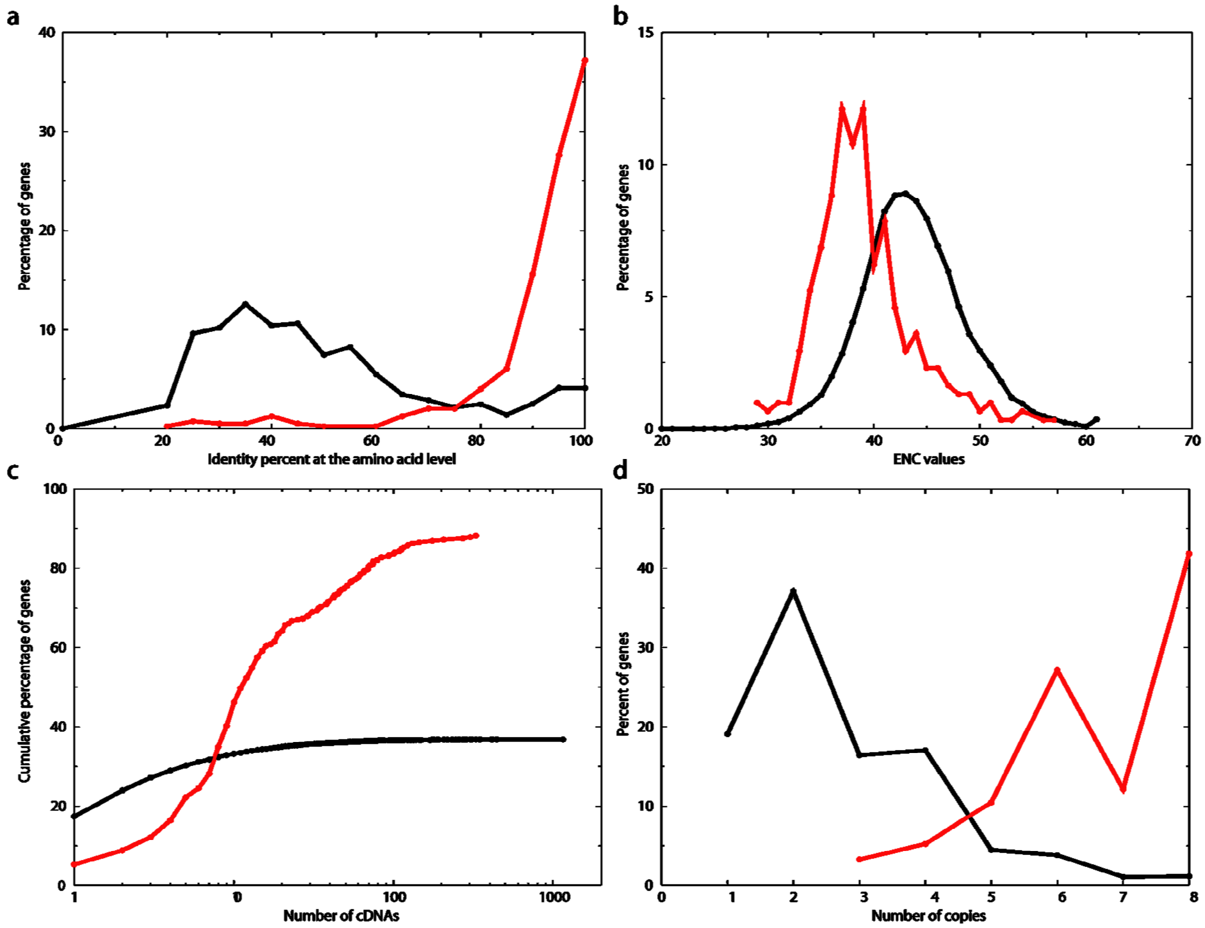


Figure S21. Characteristics of the genes with low Ks values between old duplicates. The similarity of proteins (a), the codon bias (b), the expression level (c) and the number of retained duplicates (d) are plotted for the 55 groups of paralogous genes with low nucleotide divergence (red) and for all the genes (black).

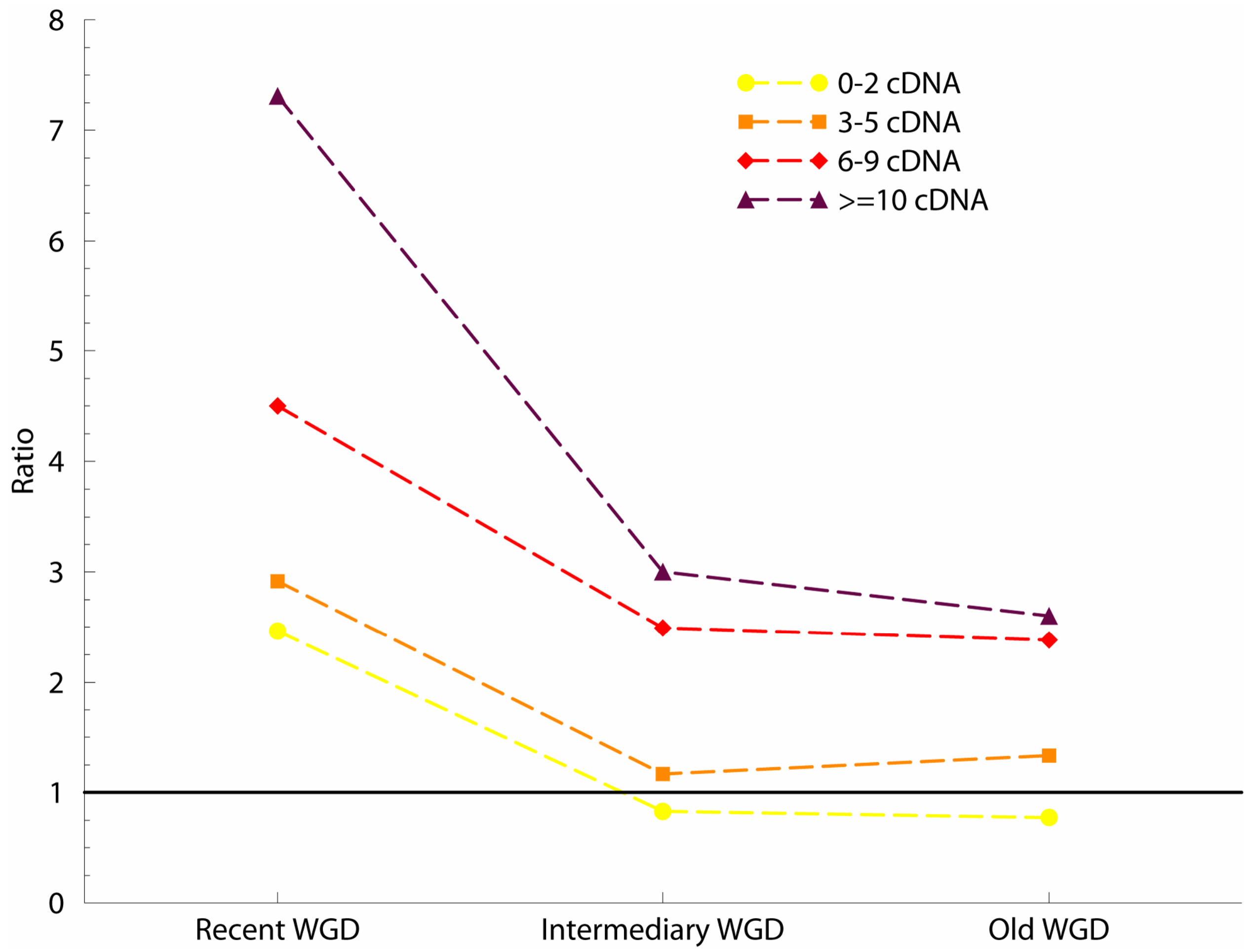


Figure S22. Expression of Paramecium proteins belonging to a complex versus retention across the three WGDs. Each point represents the ratio of duplicated genes versus non-duplicated genes, divided by the ratio of total duplicated genes versus non-duplicated genes for each WGD grouped by a defined number of ESTs matches.

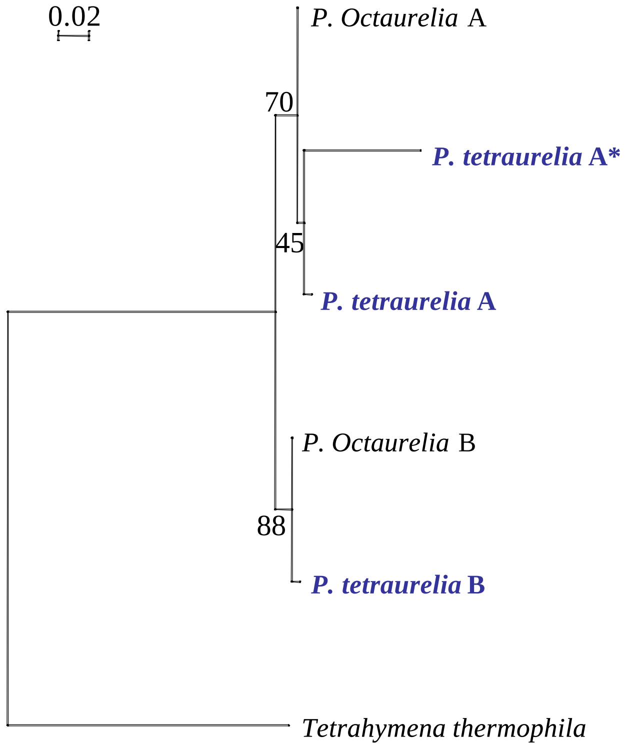


Figure S23.1. Phylogenetic tree of the Small GTPase Rab7 gene family. N=206 sites.

P. tetraurelia paralogs resulting from the recent WGD are indicated with the suffix A and B. *P. tetraurelia* A (Rab7a): GSPATP00024229001 ; *P. tetraurelia* A*: GSPATP00024226001 ; *P. tetraurelia* B (Rab7b): GSPATP00005550001 ; *P. octaurelia* A (Rab7a): AY744503 ; *P. octaurelia* B (Rab7b): AY644723 ; *T. thermophila*: AB024707. NB : Rab7a* is a recent tandem duplicate of Rab7a.

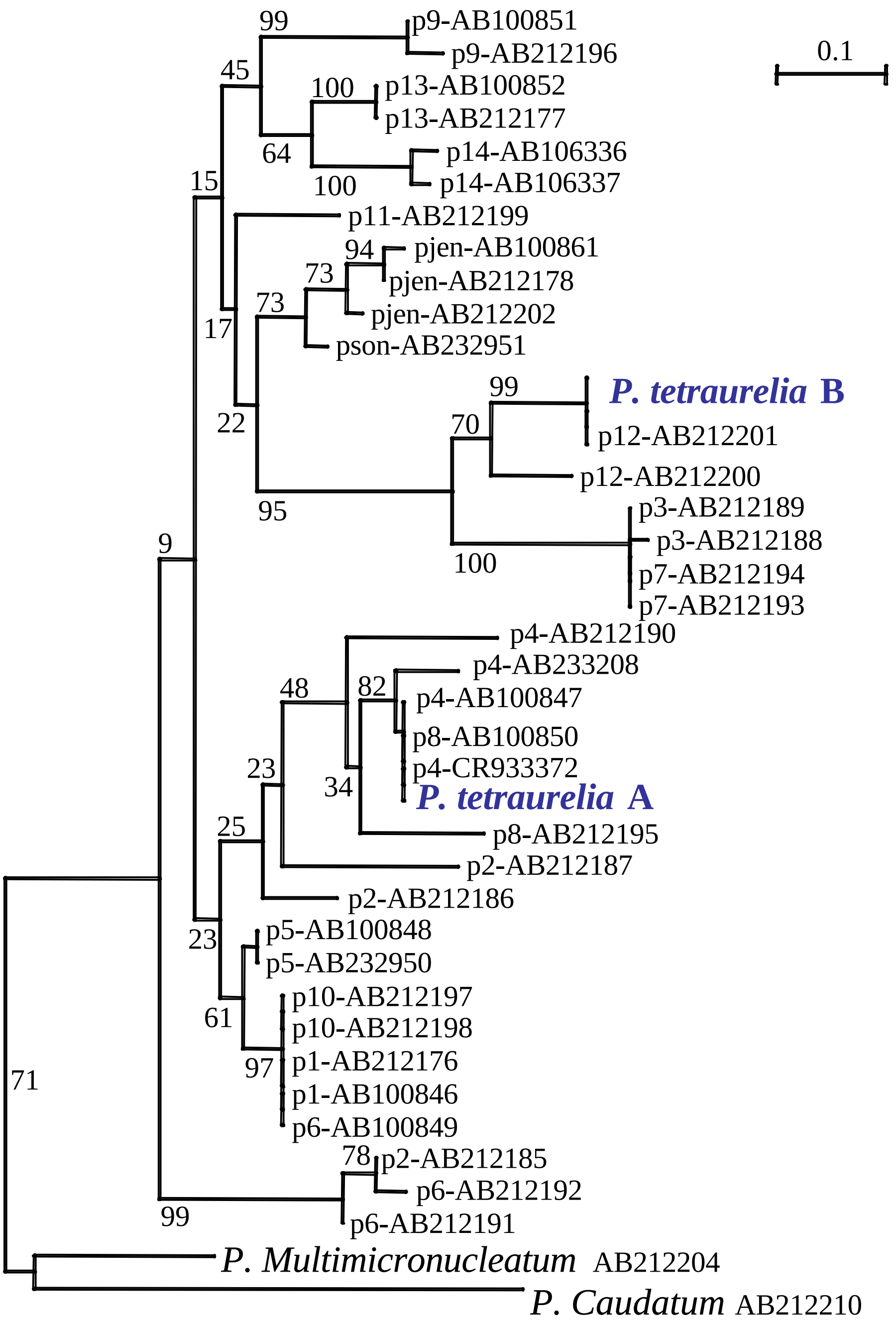


Figure S23.2a. Phylogenetic tree of the cytoplasmic HSP70 gene family. The tree was built with the Neighbor-Joining method, using synonymous distances (Ks). N=126 codons. 500 bootstrap replicates. *P. tetraurelia* paralogs resulting from the recent WGD are indicated with the suffix A and B. *P. tetraurelia* A: GSPATP00012155001 ; *P. tetraurelia* B: GSPATP00015570001. All sequences (paralogous or allelic sequences) from each species were included (accession numbers are indicated in the tree). p1 : *P. primaurelia,* p2: *P. biaurelia,* p3: *P. triaurelia,* p4: *P. tetraurelia,* p5: *P. pentaurelia,* p6: *P. sexaurelia,* p7: *P. septaurelia,* p8: *P. octaurelia,* p9: *P. novaurelia,* p10: *P. decaurelia,* p11: *P. undecaurelia,* p12: *P. dodecaurelia,* p13*: P. tredecaurelia,* p14: *P. quadecaurelia,* psonn: *P. sonneborni*, pjen: *P. jenningsi*

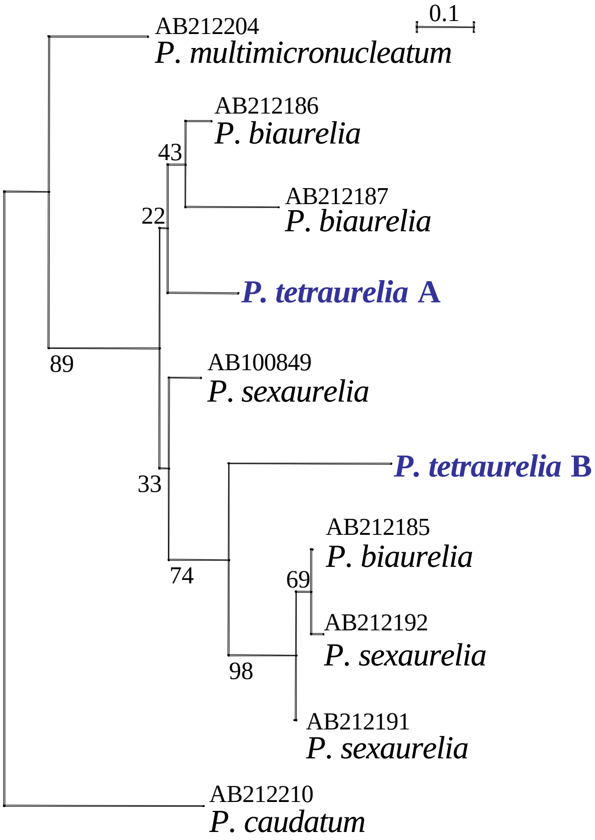


Figure S23.2b. Phylogenetic tree of the cytoplasmic HSP70 gene family : focus on *P. sexaurelia* and *P. biaurelia*. The tree was built with the Neighbor-Joining method, using synonymous distances (Ks). N=126 codons. 500 bootstrap replicates. *P. tetraurelia* paralogs resulting from the recent WGD are indicated with the suffix A and B. *P. tetraurelia* A: GSPATP00012155001 ; *P. tetraurelia* B: GSPATP00015570001. Accession numbers for other species are indicated in the tree.

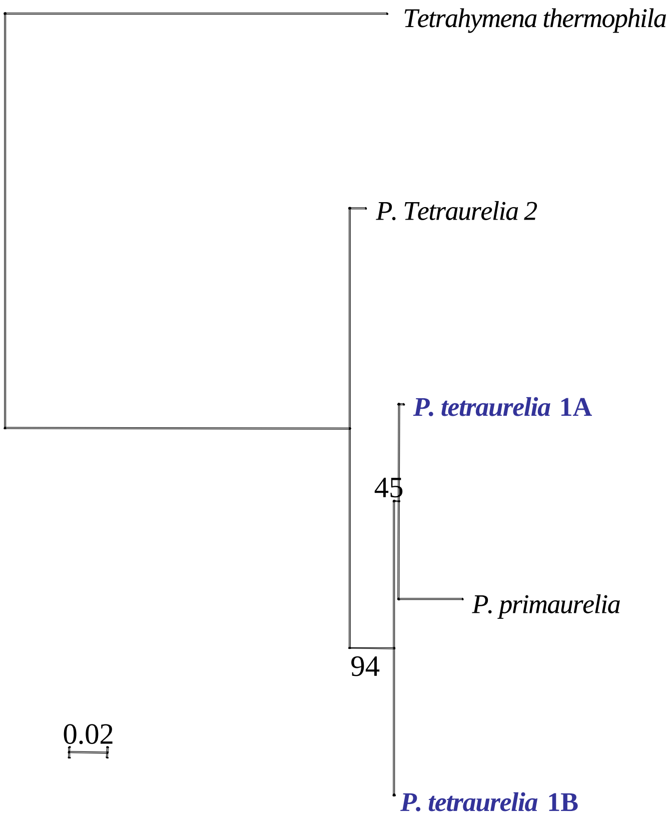


Figure S23.3. Phylogenetic tree of the phosphoglycerate kinase gene family. N=367 sites.

P. tetraurelia paralogs resulting from the recent WGD are indicated with the suffix A and B. *P. tetraurelia* 1A: GSPATP00023302001 ; *P. tetraurelia* 1B: GSPATP00021618001 ; *P. tetraurelia* 2: GSPATP00030650001 ; *P. primaurelia*: AF001849 ; *T. thermophila*: X63528

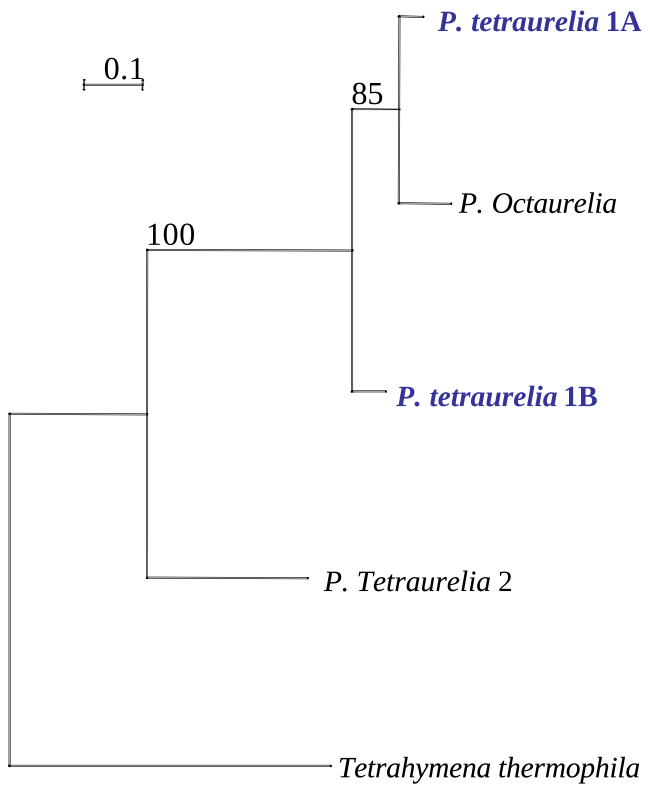


Figure S23.4. Phylogenetic tree of the Small GTPase Rab11 gene family. N=94 sites.

P. tetraurelia paralogs resulting from the recent WGD are indicated with the suffix A and B. *P. tetraurelia* 1A: GSPATP00019389001 ; *P. tetraurelia* 1B: GSPATP00017322001 ; *P. tetraurelia* 2: GSPATP00037899001 ; *P. octaurelia*: AY228707 ; *T. thermophila*: TIGR 25160

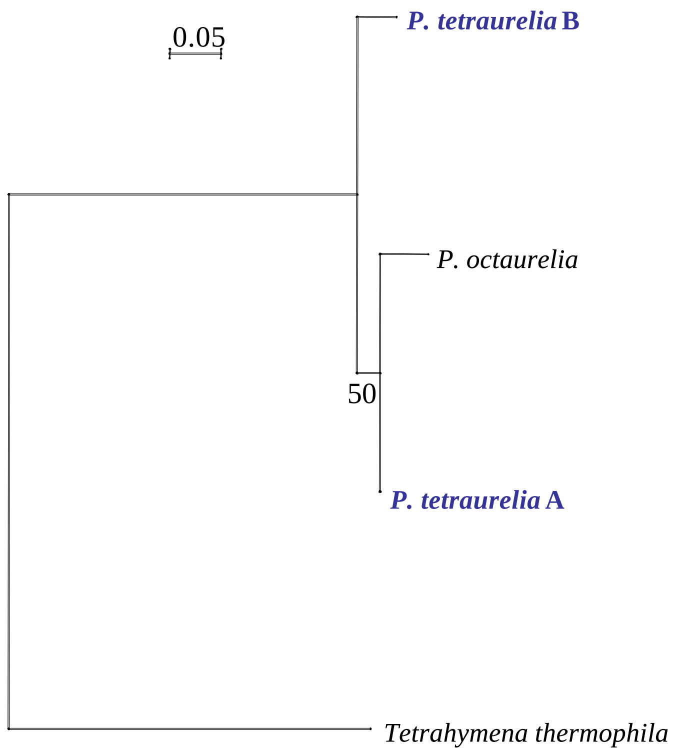


Figure S23.5. Phylogenetic tree of the GSPATP00016249001 gene family. N=137 sites. *P. tetraurelia* paralogs resulting from the recent WGD are indicated with the suffix A and B. P. octaurelia : AF346412 ; *P. tetraurelia* A: GSPATP00016249001 ; *P. tetraurelia* B: GSPATP00012521001 ; *T. thermophila*: TIGR 7050

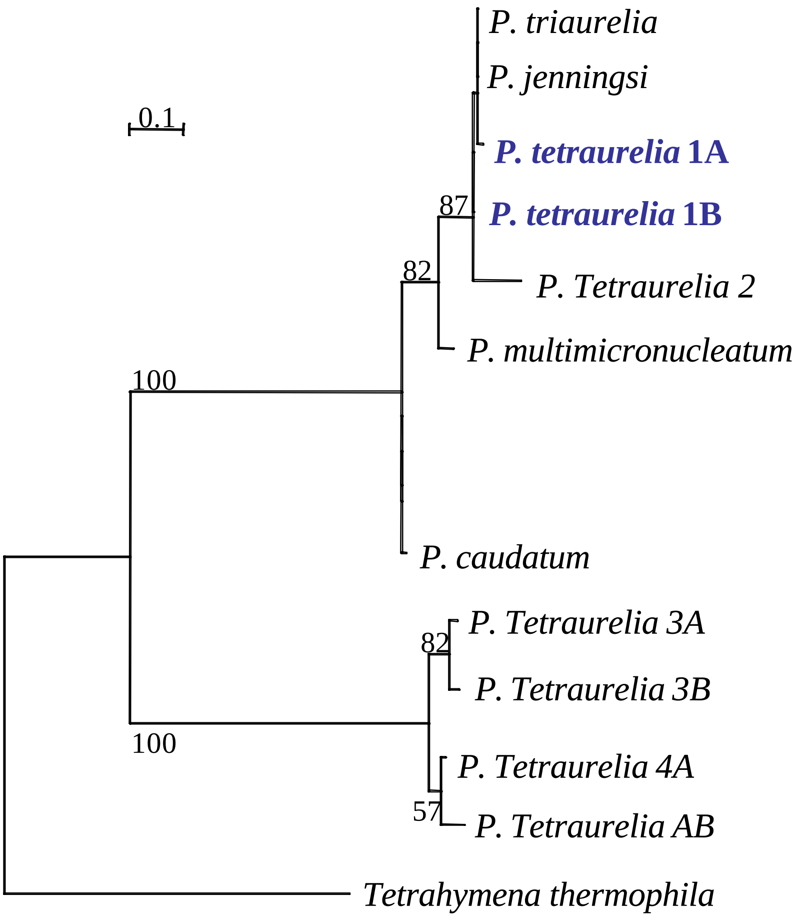


Figure S23.6. Phylogenetic tree of the hemoglobin gene family. N=116 sites

P. tetraurelia paralogs resulting from the recent WGD are indicated with the suffix A and B. *P. tetraurelia* 1A: GSPATP00012466001 ; *P. tetraurelia* 1B: GSPATP00019583001 ; *P. tetraurelia* 2: GSPATP00010635001 ; *P. tetraurelia* 3A: GSPATP00028689001 ; *P. tetraurelia* 3B: GSPATP00026932001 ; *P. tetraurelia* 4A: GSPATP00028853001 ; *P. tetraurelia* 4B: GSPATP00035893001 ; *P. triaurelia*: D49688 ; *P. jenningsi*: D49689 *;* *P. multimicronucleatum*: D49687 ; *P. caudatum*: M99047 ; *T. thermophila*: D13919

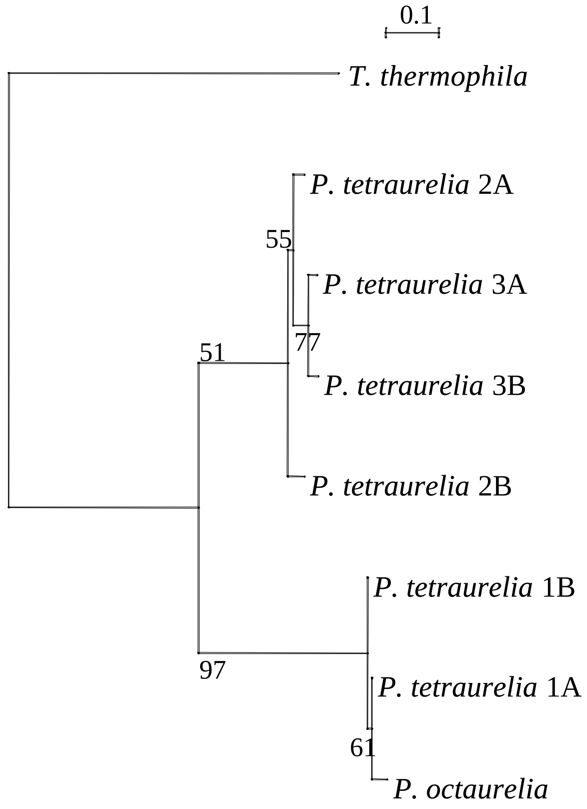


Figure S23.7. Phylogenetic tree of the MPK2 serine/threonine protein-like kinase gene family. N=122 sites

P. tetraurelia paralogs resulting from the recent WGD are indicated with the suffix A and B. *P. tetraurelia* 1A: GSPATP00025481001 ; *P. tetraurelia* 1B: GSPATP00031653001 ; *P. tetraurelia* 2A: GSPATP00031569001 ; *P. tetraurelia* 2B: GSPATP00025382001 *P. tetraurelia* 3A: GSPATP00027980001 ; *P. tetraurelia* 3B: GSPATP00014044001 *P. octaurelia*: AF346410 ; *T. thermophila*: AY426250

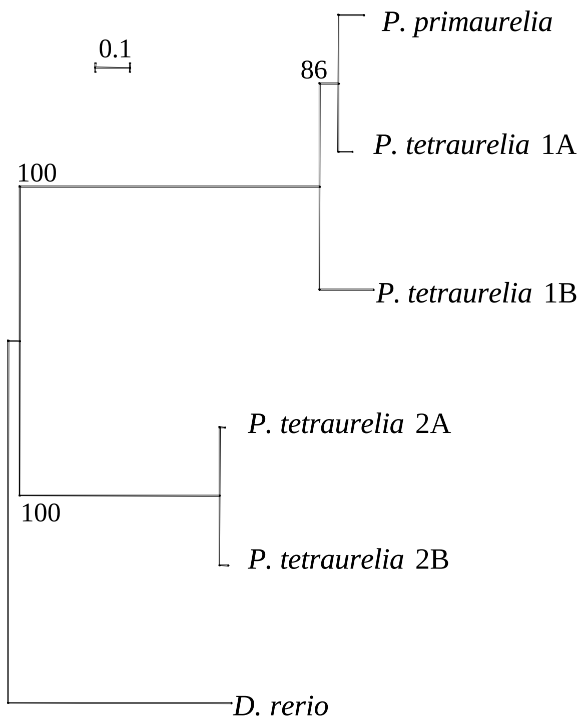


Figure S23.8. Phylogenetic tree of the NMP kinase gene family. N=195 sites

P. tetraurelia paralogs resulting from the recent WGD are indicated with the suffix A and B. *P. tetraurelia* 1A: GSPATP00017197001 ; *P. tetraurelia* 1B: GSPATP00018531001 ; *P. tetraurelia* 2A: GSPATP00004957001 ; *P. tetraurelia* 2B: GSPATP00003275001 ; *P. primaurelia*: Y13117 / CAA73579 / O00846 ; *Danio rerio*: BC049446

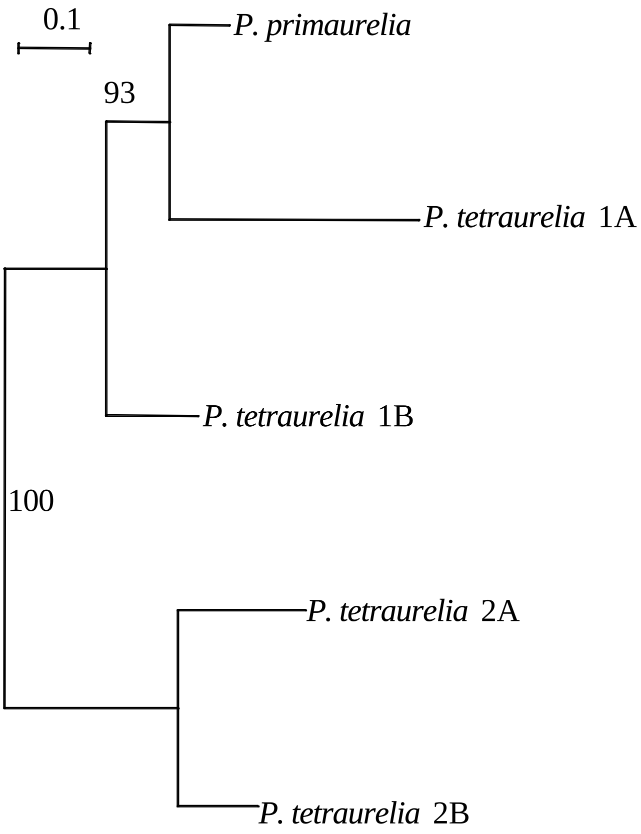


Figure S23.9. Phylogenetic tree of the GSPATP00018530001 gene family. N=148 sites

P. tetraurelia paralogs resulting from the recent WGD are indicated with the suffix A and B. *P. tetraurelia* 1A: GSPATP00017194001 ; *P. tetraurelia* 1B: GSPATP00018530001 ; *P. tetraurelia* 2A: GSPATP00029311001 ; *P. tetraurelia* 2B: GSPATP00024732001 ; *P. primaurelia*: Y13117 / CAA73577 / O00844.

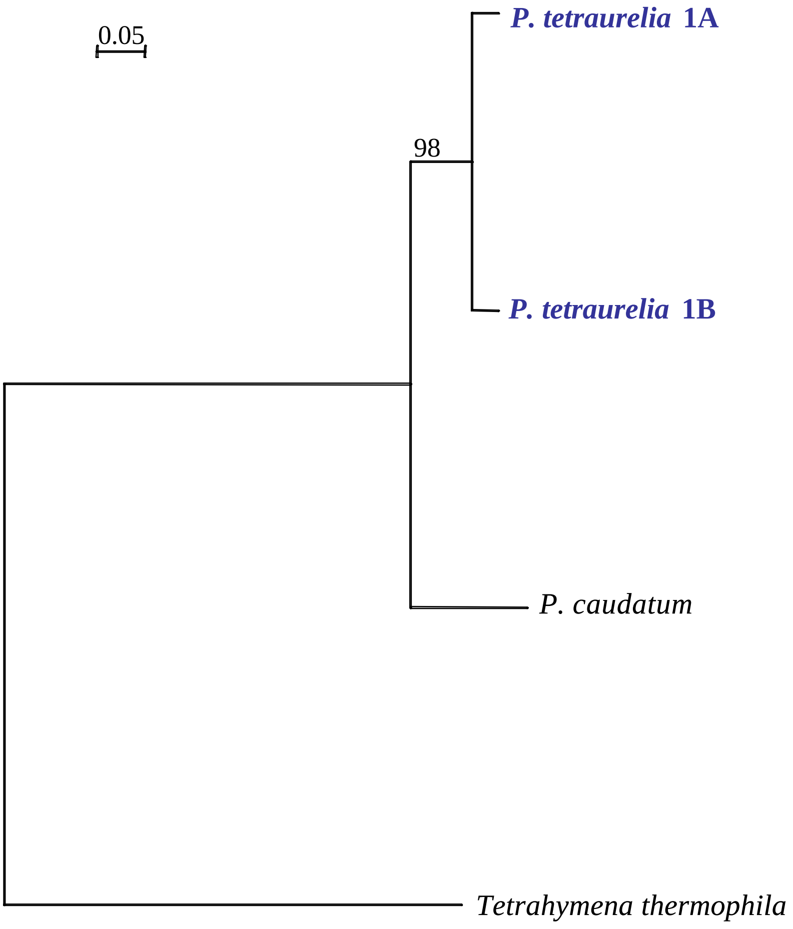


Figure S23.10. Phylogenetic tree of the dad-1 (Defender against cell death) gene family. N=118 sites

P. tetraurelia paralogs resulting from the recent WGD are indicated with the suffix A and B. *P. tetraurelia* 1A: GSPATP0002402500 ; *P. tetraurelia* 1B: GSPATP00026546001 ; *P. caudatum*: AB175335 ; *T. thermophila*: TIGR 21571

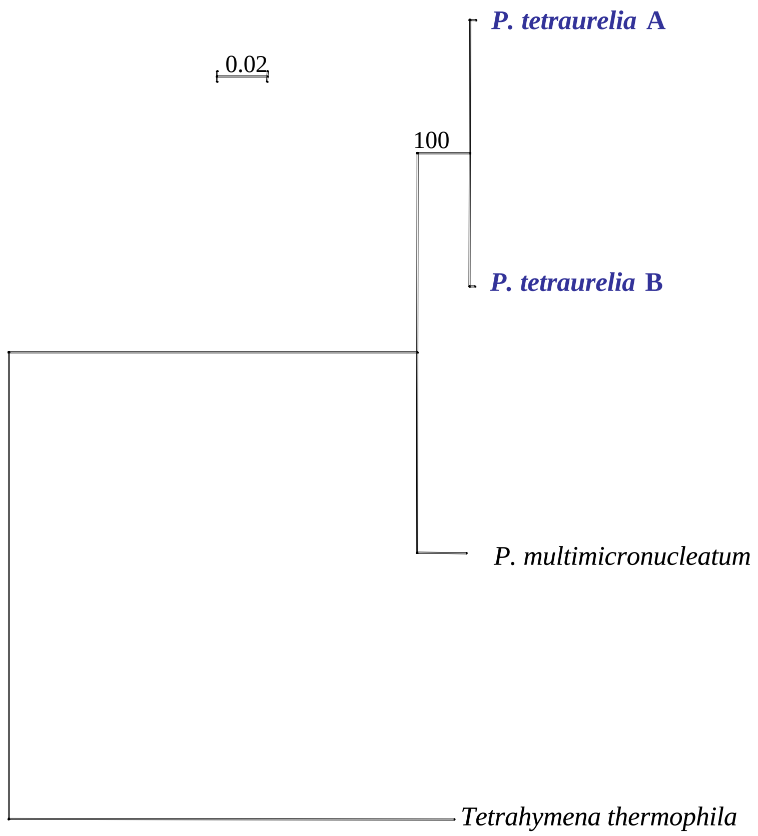


Figure S23.11. Phylogenetic tree of the enolase gene family. N=356 sites

P. tetraurelia paralogs resulting from the recent WGD are indicated with the suffix A and B. *P. tetraurelia* A: GSPATP00019191001 ; *P. tetraurelia* B: GSPATP00004487001 *; P. multimicronucleatum*: AF348926 ; *T. thermophila*: TIGR 8254659

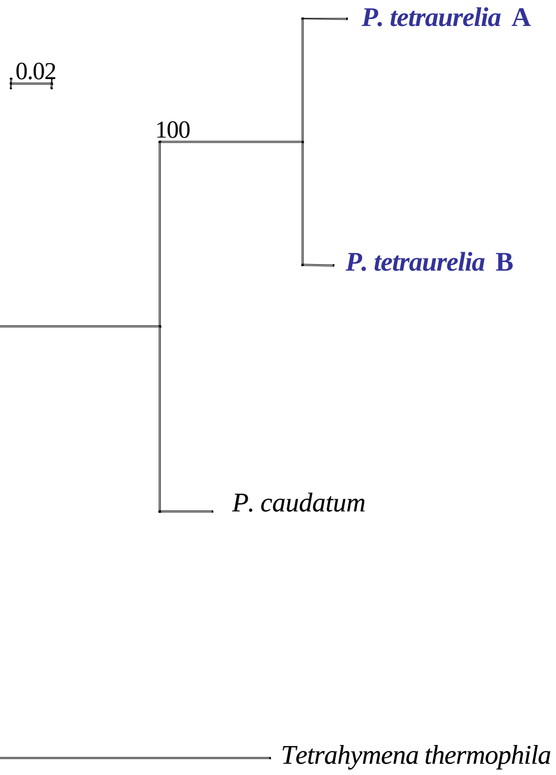


Figure S23.12. Phylogenetic tree of the HSP60 gene family. N=172 sites

P. tetraurelia paralogs resulting from the recent WGD are indicated with the suffix A and B. *P. tetraurelia* 1A: GSPATP00015732001 ; *P. tetraurelia* 1B: GSPATP00027766001 ; *P. caudatum*: AB048353 ; *T. thermophila*: TIGR 8254617

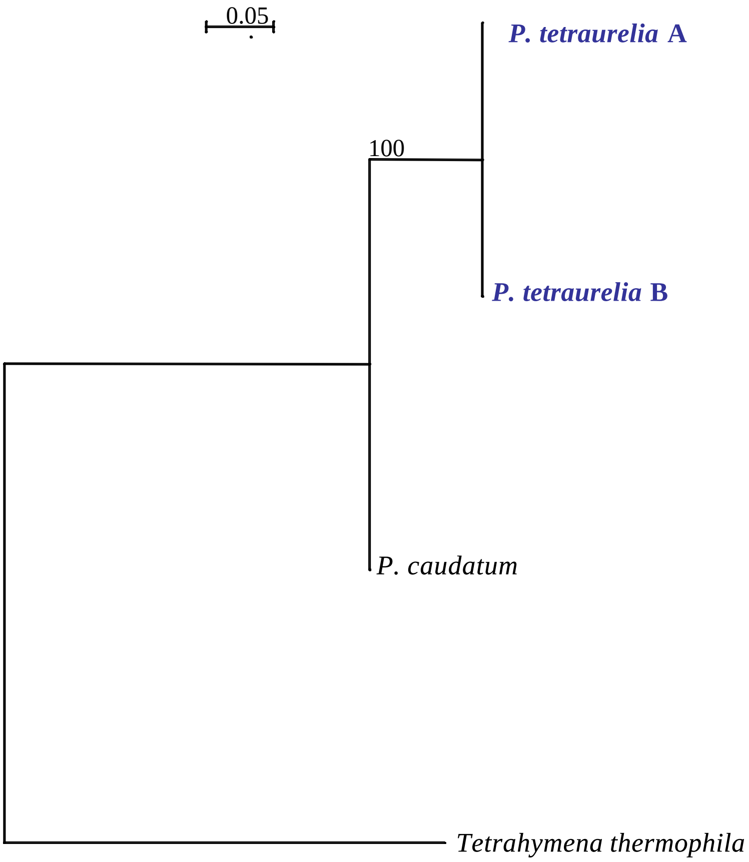


Figure S23.13. Phylogenetic tree of the ribosomal protein L4 gene family. N=179 sites

P. tetraurelia paralogs resulting from the recent WGD are indicated with the suffix A and B. *P. tetraurelia* A: GSPATP00023827001 ; *P. tetraurelia* B: GSPATP00033610001 ; *P. caudatum*: AB071329 ; *T. thermophila*: TIGR 8254610

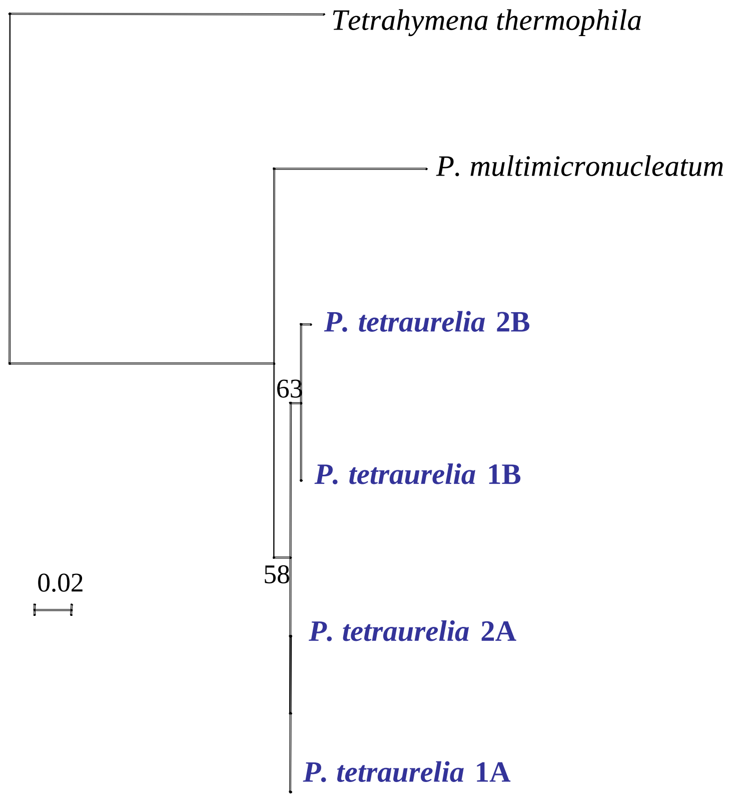


Figure S23.14. Phylogenetic tree of the vacuolar ATPase beta gene family. N=198 sites

P. tetraurelia paralogs resulting from the recent WGD are indicated with the suffix A and B. *P. tetraurelia* 1A: GSPATP00006101001 ; *P. tetraurelia* 1B: GSPATP00014624001 ; *P. tetraurelia* 2A: GSPATP00010991001 ; *P. tetraurelia* 2B: GSPATP00014470001 *P. multimicronucleatum*: AB066280 ; *T. thermophila*: 4655

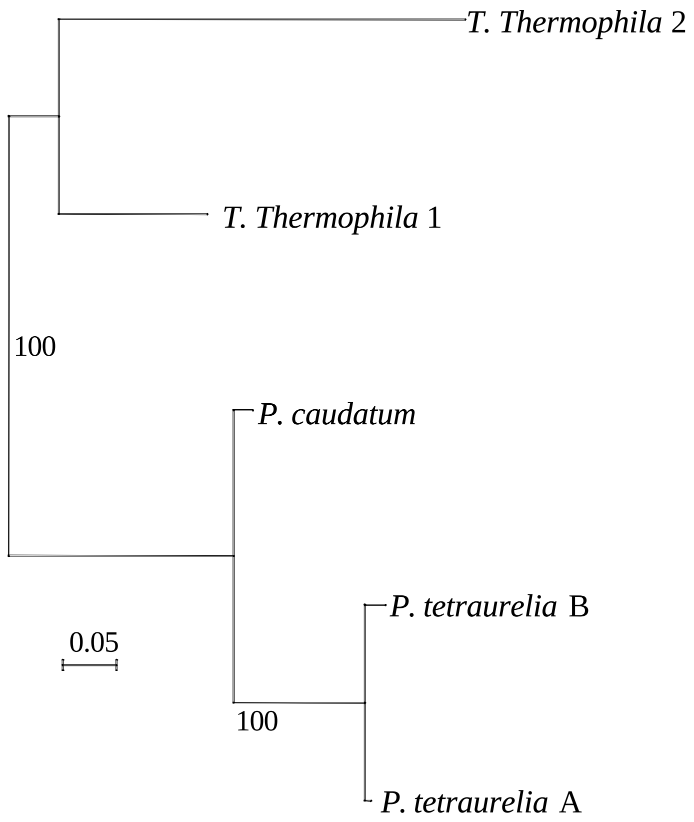


Figure S23.15. Phylogenetic tree of the pcstk-1 MAP kinase gene family. N=295 sites

P. tetraurelia paralogs resulting from the recent WGD are indicated with the suffix A and B. *P. tetraurelia* 1A: GSPATP00021256001 ; *P. tetraurelia* 1B: GSPATP00038202001 ; *P. caudatum*: AB195231 : *T. thermophila* 1: TIGR 29899 ; *T. thermophila* 2: TIGR 10873

Supplementary Tables

Table S1. Sequencing statistics.

Insert sizes	Vector type	Number of reads	Trimmed read lengths (b)	Fraction paired (%)	Genomic coverage*
3.1 kb	HC plasmid	334 705	816	94.4	3.64x
8.6 kb	LC plasmid	191 288	775	92.5	1.98x
2.5 kb	LC plasmid	159 463	713	91.4	1.52x
9.7 kb	LC plasmid	94 648	771	92.8	0.97x
6.7 kb	LC plasmid	151 625	844	93.9	1.71x
6.6 kb	LC plasmid	219 386	831	94.6	2.43x
17 kb	BAC	83 887	697	96.4	0.78x
total		1 235 002	791	93.7	13.03x

Table S2. Macronuclear chromosome status. The table presents data for the 188 scaffolds larger than 45 kb, representing 96% of the assembly. For each scaffold, the columns from left to right present the scaffold name, size, status, presence of left and right telomere repeats, number of sequence gaps, total estimated size of the sequence gaps and the fraction of the scaffold that is contained in the gaps. The status column summarizes the evidence that this scaffold is a complete macronuclear chromosome: 'complete' if both left and right telomeres where found, 'incomplete' if they were not, and 'incomplete*' if the alignment of pairs of scaffolds related by the recent WGD indicates that the scaffold does extend into the sub-telomeric region even though the telomere repeats themselves were not detected.

Telomere repeats were detected in two ways. First, since telomere repeats were masked during assembly, we re-mapped telomere reads to the assembly. The telomere read library consisted of 15,242 high quality sequence reads containing at least three repeats of $C_3(AC)A_2$ hexamers[37], with at most one mismatch. These reads were mapped to the assembly by megablast. The Bioperl library[38] was used to store the data in a relational database and produce graphical representations of the scaffolds and match positions. Scaffolds with correctly oriented telomere reads clustered at their ends were considered to have a telomere ('x' in the table). Second, a small number of scaffolds had no terminal re-mapped telomere reads but the sequence did have recognizable 5' or 3' telomere repeats. These scaffolds were also scored positively ('z' in the table).

scaffold	size (bp)	status	telomeres		gaps number	size (bp)	fraction
			left	right			
scaffold_1	981684	complete	X	X	4	1073	0.0011
scaffold_2	950118	complete	X	X	5	2225	0.0023
scaffold_3	893859	complete	X	X	10	1432	0.0016
scaffold_4	817033	incomplete	X		8	4276	0.0052
scaffold_5	770190	complete	Z	X	8	4761	0.0062
scaffold_6	752274	complete	X	X	5	815	0.0011
scaffold_7	741107	complete	X	X	5	995	0.0013
scaffold_8	739885	complete	X	X	7	2223	0.0030
scaffold_9	703586	complete	X	X	4	1405	0.0020
scaffold_10	685516	complete	X	X	2	1957	0.0029
scaffold_11	682700	incomplete*	X		3	485	0.0007
scaffold_12	675213	incomplete		X	1	150	0.0002
scaffold_13	667686	complete	X	X	5	964	0.0014
scaffold_14	658971	complete	X	X	1	270	0.0004
scaffold_15	651731	complete	X	X	4	1594	0.0024
scaffold_16	645506	incomplete*	X		4	462	0.0007
scaffold_17	641603	complete	Z	X	8	3718	0.0058
scaffold_18	604692	complete	X	X	6	910	0.0015
scaffold_19	601676	complete	X	X	4	1855	0.0031
scaffold_20	584889	incomplete*	X		5	2935	0.0050
scaffold_21	573080	complete	X	X	4	1213	0.0021
scaffold_22	570993	complete	X	X	7	828	0.0015
scaffold_23	563831	complete	X	X	3	466	0.0008
scaffold_24	562444	complete	X	X	5	7197	0.0128
scaffold_25	554326	complete	X	X	5	737	0.0013
scaffold_26	553472	complete	X	X	8	1165	0.0021
scaffold_27	545223	incomplete*		X	3	450	0.0008
scaffold_28	539671	complete	X	X	5	1004	0.0019
scaffold_29	537272	incomplete	X		4	1302	0.0024
scaffold_30	534934	incomplete			2	111	0.0002
scaffold_31	528592	incomplete		Z	3	1939	0.0037
scaffold_32	527326	complete	X	X	5	743	0.0014
scaffold_33	522731	complete	X	X	0	0	0.0000
scaffold_34	509599	incomplete*	X		4	763	0.0015
scaffold_35	509102	incomplete*			1	150	0.0003

scaffold_36	507435	incomplete*	X		5	831	0.0016
scaffold_37	505928	complete	X	X	3	498	0.0010
scaffold_38	502993	complete	X	X	4	981	0.0020
scaffold_39	495371	complete	X	X	0	0	0.0000
scaffold_40	494621	complete	X	X	5	1217	0.0025
scaffold_41	483735	incomplete		Z	0	0	0.0000
scaffold_42	483623	complete	X	X	4	449	0.0009
scaffold_43	481750	incomplete			4	1565	0.0032
scaffold_44	479897	complete	X	X	5	933	0.0019
scaffold_45	476979	complete	X	X	3	532	0.0011
scaffold_46	475397	incomplete			2	521	0.0011
scaffold_47	473319	complete	X	X	14	8674	0.0183
scaffold_48	468060	complete	X	X	3	4293	0.0092
scaffold_49	465594	incomplete		X	3	870	0.0019
scaffold_50	463635	complete	X	X	3	450	0.0010
scaffold_51	463379	complete	X	X	3	5474	0.0118
scaffold_52	461155	complete	X	X	4	632	0.0014
scaffold_53	459636	incomplete	X		9	2355	0.0051
scaffold_54	459346	complete	X	X	1	150	0.0003
scaffold_55	458689	incomplete		X	3	546	0.0012
scaffold_56	456850	complete	X	X	3	383	0.0008
scaffold_57	454313	complete	X	X	1	150	0.0003
scaffold_58	446585	complete	X	X	9	2815	0.0063
scaffold_59	440049	incomplete			1	1556	0.0035
scaffold_60	437965	complete	X	X	4	3001	0.0069
scaffold_61	435734	complete	X	X	2	197	0.0005
scaffold_62	425387	incomplete	X		4	933	0.0022
scaffold_63	421209	complete	X	X	1	150	0.0004
scaffold_64	413286	complete	X	X	3	402	0.0010
scaffold_65	410769	complete	X	X	1	150	0.0004
scaffold_66	407364	complete	X	X	1	150	0.0004
scaffold_67	405785	complete	X	X	2	3666	0.0090
scaffold_68	405512	complete	X	X	6	1422	0.0035
scaffold_69	390565	complete	X	X	5	4240	0.0109
scaffold_70	388433	complete	X	X	3	1282	0.0033
scaffold_71	387862	complete	X	X	5	8460	0.0218
scaffold_72	386812	incomplete*		X	5	1399	0.0036
scaffold_73	371967	incomplete			1	257	0.0007
scaffold_74	368269	complete	X	Z	4	733	0.0020
scaffold_75	363622	complete	X	X	3	237	0.0007
scaffold_76	362286	incomplete	X		5	1040	0.0029
scaffold_77	360245	complete	X	X	5	2764	0.0077
scaffold_78	358839	complete	X	X	1	150	0.0004
scaffold_79	358756	complete	X	X	1	349	0.0010
scaffold_80	358292	complete	X	X	4	2048	0.0057
scaffold_81	357618	complete	X	X	6	2095	0.0059
scaffold_82	357438	incomplete	X		3	405	0.0011
scaffold_83	355138	complete	X	X	2	818	0.0023
scaffold_84	354111	complete	X	X	4	623	0.0018
scaffold_85	349785	complete	X	X	0	0	0.0000
scaffold_86	349122	complete	X	X	5	1255	0.0036
scaffold_87	342910	incomplete			1	102	0.0003
scaffold_88	342040	complete	X	X	7	917	0.0027
scaffold_89	341488	incomplete	X		1	150	0.0004
scaffold_90	339195	complete	X	X	2	2623	0.0077
scaffold_91	338364	complete	X	X	1	2766	0.0082
scaffold_92	332550	complete	X	X	3	452	0.0014
scaffold_93	332513	complete	X	X	4	1587	0.0048

scaffold_94	330718	complete	X	X	1	1007	0.0030
scaffold_95	330322	complete	X	X	1	150	0.0005
scaffold_96	328795	complete	X	X	1	199	0.0006
scaffold_97	323246	complete	X	X	3	361	0.0011
scaffold_98	319182	complete	X	X	2	1943	0.0061
scaffold_99	315007	complete	X	X	6	4177	0.0133
scaffold_100	312801	complete	X	X	0	0	0.0000
scaffold_101	310408	incomplete	X		1	2459	0.0079
scaffold_102	310338	complete	X	X	6	2845	0.0092
scaffold_103	306506	complete	X	X	1	150	0.0005
scaffold_104	304222	complete	X	X	0	0	0.0000
scaffold_105	303662	incomplete*	X		4	456	0.0015
scaffold_106	302836	complete	X	X	1	150	0.0005
scaffold_107	302599	complete	X	X	2	284	0.0009
scaffold_108	302230	complete	X	X	3	3831	0.0127
scaffold_109	302220	incomplete*			2	1086	0.0036
scaffold_110	301577	incomplete*		X	2	312	0.0010
scaffold_111	300637	complete	X	X	4	515	0.0017
scaffold_112	297279	incomplete		X	2	3892	0.0131
scaffold_113	295349	complete	X	X	5	750	0.0025
scaffold_114	293094	incomplete*	X		1	150	0.0005
scaffold_115	292917	incomplete*	X		1	387	0.0013
scaffold_116	290952	incomplete*	X		2	300	0.0010
scaffold_117	289312	complete	X	X	4	452	0.0016
scaffold_118	285946	complete	X	X	3	861	0.0030
scaffold_119	284277	incomplete	Z		5	710	0.0025
scaffold_120	280595	incomplete*	X		4	2279	0.0081
scaffold_121	275544	complete	X	X	1	150	0.0005
scaffold_122	275394	complete	X	X	3	585	0.0021
scaffold_123	274475	complete	X	X	3	450	0.0016
scaffold_124	274028	incomplete*	X		6	1317	0.0048
scaffold_125	271968	incomplete			1	56	0.0002
scaffold_126	271500	complete	X	X	2	300	0.0011
scaffold_127	270377	complete	X	X	4	813	0.0030
scaffold_128	268562	incomplete		X	1	84	0.0003
scaffold_129	266443	complete	X	X	2	1470	0.0055
scaffold_130	264698	complete	X	X	1	719	0.0027
scaffold_131	263758	incomplete*	X		1	143	0.0005
scaffold_132	262693	complete	X	X	5	1699	0.0065
scaffold_133	262487	complete	X	X	5	750	0.0029
scaffold_134	262358	complete	X	X	0	0	0.0000
scaffold_135	261594	incomplete			3	450	0.0017
scaffold_136	260901	incomplete	X		2	300	0.0011
scaffold_137	260781	incomplete			1	150	0.0006
scaffold_138	256454	incomplete*	Z		4	6410	0.0250
scaffold_139	255169	incomplete	X		0	0	0.0000
scaffold_140	252200	incomplete*	X		0	0	0.0000
scaffold_141	247167	incomplete*		X	4	3615	0.0146
scaffold_142	247095	complete	X	X	4	1017	0.0041
scaffold_143	247008	incomplete			9	1710	0.0069
scaffold_144	245749	complete	X	X	1	525	0.0021
scaffold_145	245673	complete	X	X	2	1172	0.0048
scaffold_146	244492	complete	X	X	2	300	0.0012
scaffold_147	242120	incomplete	X		0	0	0.0000
scaffold_148	236632	incomplete	X		2	300	0.0013
scaffold_149	235637	complete	X	X	1	746	0.0032
scaffold_150	235572	complete	X	X	2	223	0.0009
scaffold_151	234462	incomplete	X		3	450	0.0019

scaffold_152	227017	complete	X	X	4	4215	0.0186
scaffold_153	226181	incomplete	X		3	1081	0.0048
scaffold_154	225099	incomplete		X	3	450	0.0020
scaffold_155	224907	complete	X	X	2	152	0.0007
scaffold_156	223177	incomplete*		X	8	1916	0.0086
scaffold_157	222233	complete	X	X	4	6541	0.0294
scaffold_158	221018	complete	X	X	3	450	0.0020
scaffold_159	220852	complete	X	X	4	600	0.0027
scaffold_556	215862	complete	X	X	1	1115	0.0052
scaffold_160	215836	complete	X	X	4	640	0.0030
scaffold_161	215326	complete	X	X	2	913	0.0042
scaffold_162	213900	complete	X	X	2	2433	0.0114
scaffold_163	212828	complete	X	X	4	1659	0.0078
scaffold_164	212585	incomplete	X		3	716	0.0034
scaffold_165	212273	complete	X	X	1	904	0.0043
scaffold_166	202677	complete	X	X	3	764	0.0038
scaffold_167	199358	complete	X	X	0	0	0.0000
scaffold_168	191157	incomplete			4	1555	0.0081
scaffold_169	178157	complete	X	X	1	824	0.0046
scaffold_170	175797	incomplete			3	1443	0.0082
scaffold_171	174740	incomplete		X	2	381	0.0022
scaffold_172	146481	complete	X	X	1	150	0.0010
scaffold_173	144165	incomplete	X		0	0	0.0000
scaffold_174	143093	incomplete		X	1	66	0.0005
scaffold_175	139747	incomplete		X	1	76	0.0005
scaffold_176	138878	incomplete			5	698	0.0050
scaffold_177	138786	incomplete	X		1	2088	0.0150
scaffold_178	138406	incomplete*		X	2	2174	0.0157
scaffold_179	100681	incomplete			2	1270	0.0126
scaffold_180	92100	incomplete		X	0	0	0.0000
scaffold_181	90629	incomplete	X		1	287	0.0032
scaffold_182	84216	incomplete		X	0	0	0.0000
scaffold_183	81199	incomplete			1	222	0.0027
scaffold_184	72531	incomplete		X	2	164	0.0023
scaffold_185	58408	incomplete		X	0	0	0.0000
scaffold_186	51533	incomplete		X	0	0	0.0000
scaffold_187	47800	incomplete	X		1	150	0.0031

	protist			plant		vertebrate						fungus				invertebrate			
	P.t.	P.f.	G.t.	O.s.	A.t.	T.n.	T.r.	D.r.	G.g.	C.f.	M.m.	S.p.	S.c.	E.g.	E.c.	C.e.	A.g.	D.m.	A.m.
AA	1.04	0.95	1.12	1.16	1.14	1.18	1.15	1.11	1.13	1.13	1.08	1.17	1.13	1.03	1.12	1.31	1.23	1.23	1.16
AC	0.77	1.00	0.80	0.86	0.93	0.93	0.93	0.97	0.85	0.79	0.87	0.95	0.90	0.93	0.77	0.83	0.96	0.86	0.80
AG	1.00	0.75	0.93	0.93	0.98	1.06	1.04	0.98	1.19	1.17	1.21	0.96	0.99	1.00	1.07	0.93	0.85	0.87	0.80
AT	1.02	1.09	0.95	1.01	0.90	0.83	0.88	0.92	0.84	0.89	0.86	0.88	0.93	1.03	0.99	0.83	0.92	0.95	1.03
CA	1.07	1.10	0.91	1.06	1.08	1.23	1.24	1.26	1.27	1.15	1.23	1.05	1.03	1.07	1.07	1.05	1.12	1.12	0.86
CC	1.23	1.84	1.40	1.13	1.08	1.04	1.05	1.04	1.10	1.29	1.21	1.01	1.07	0.92	1.21	1.10	0.98	1.07	1.07
CG	**0.52**	**0.91**	**1.07**	**0.88**	**0.78**	**0.60**	**0.60**	**0.53**	**0.25**	**0.26**	**0.19**	**0.96**	**0.92**	**1.02**	**0.58**	**0.96**	**1.06**	**0.94**	**1.67**
CT	1.00	0.78	0.89	0.94	0.99	1.06	1.04	0.98	1.20	1.17	1.21	0.96	0.97	0.98	1.09	0.92	0.85	0.87	0.79
GA	1.13	0.83	0.98	0.99	1.07	0.98	1.00	0.90	0.98	1.03	1.03	0.97	0.98	0.88	1.07	1.13	0.94	0.88	1.15
GC	1.12	1.19	1.12	1.06	0.90	1.08	1.05	1.18	1.13	0.97	0.93	1.19	1.14	1.27	1.00	0.97	1.15	1.32	1.07
GG	1.23	1.93	1.43	1.14	1.08	1.03	1.04	1.04	1.09	1.28	1.20	1.01	1.08	0.92	1.18	1.09	0.97	1.06	1.06
GT	0.77	0.98	0.79	0.85	0.94	0.93	0.94	0.97	0.86	0.80	0.88	0.94	0.90	0.92	0.76	0.84	0.97	0.86	0.79
TA	0.90	1.06	0.89	0.80	0.78	0.66	0.67	0.80	0.70	0.74	0.74	0.83	0.87	1.01	0.74	0.61	0.72	0.76	0.84
TC	1.13	0.83	0.97	0.99	1.08	0.97	0.99	0.90	0.98	1.02	1.02	0.96	0.98	0.88	1.04	1.12	0.94	0.88	1.14
TG	1.06	1.11	0.86	1.06	1.09	1.23	1.24	1.26	1.27	1.15	1.23	1.05	1.00	1.05	1.10	1.04	1.13	1.13	0.86
TT	1.04	0.95	1.15	1.16	1.13	1.18	1.14	1.11	1.11	1.13	1.07	1.16	1.14	1.05	1.12	1.29	1.23	1.23	1.16

Table S3. Dinucleotide frequencies. The observed/expected dinucleotide frequencies of 19 genomes were calculated, using the intergenic regions in order to avoid bias from codon usage. P.t. Paramecium tetraurelia; P.f. Plasmodium falciparum; G.t. Guillardia theta; O.s. Oryza sativa; A.t. Arabidopsis thaliana; T.n. Tetraodon nigroviridis; T.r. Takifugu rubripes; D.r. Danio rerio; G.g. Gallus gallus; C.f. Canis familiaris; M.m. Mus musculus; S.p. Schizosaccharomyces pombe; S.c. Saccharomyces cerevisiae; E.g. Eremothecium gossypii; E.c. Encephalitozoon cuniculi; C.e. Caenorhabditis elegans; A.g. Anopheles gambiae; D.m. Drosophila melanogaster; A.m. Apis mellifera

CpG depression is found in the *Paramecium*, plant and vertebrate genomes. It is generally accepted that this depletion, in plants and vertebrates, is a consequence of methylation at CpG sites, because methylation promotes the mutation of cytosines[39]. However, no cytosine methylation has ever been observed in *Paramecium* and methods that detect a cytosine methyltransferase gene in the genomes of *Plasmodium* and *Dictyostelium* as well as those of plants and animals[40] fail to detect any homolog in the *P. tetraurelia* macronuclear draft genome (Loïc Ponger, personal communication). A similar situation, CpG depression in the apparent absence of a cytosine methyltransferase gene, is also found in the genome of the microsporidian *E. cuniculi*[40].

Table S4. cDNA sequencing statistics.

Library name	Differentiation stage	Valid sequences	Trimmed read length (bp)
LK0AAA	Vegetative cells, 35°C	9634	726
LK0ABA	Vegetative cells, 39°C	8821	714
LK0ACA	Vegetative cells, 27°C	25602	733
LK0ADA	Beginning of meiosis	10037	781
LK0AEA	Meiosis and beginning of macronuclear development	18551	689
LK0AFA	Autogamy	17353	503
total	All	89998	682

Table S5. Annotation confidence. Statistical values on different gene sets confirmed by only one kind of evidence. 'All predictions' line represents the global Genoscope gene set.

Gene category	Number	Size	Exon size	Exons/gene
1. *Ab initio* only	763	567	305	1.79
2. *Paramecium* evidence only	7,107	760	300	2.55
3. Alveolata evidence only	1,026	847	300	2.02
4. Uniprot match	19,220	1,767	437	3.85
5. Mixed evidence	11,730	1,394	437	3.07
All predictions	39,642	1,431	419	3.28

Table S6. Orthologous genes with ten species.

Species	Total number of orthologous genes	Average %ID at amino acid level
Paramecium – Human	1,671	39.32
Paramecium – Drosophila	1,442	39.22
Paramecium – Arabidopsis	1,505	40.04
Paramecium – Plasmodium	898	39.38
Paramecium – Neurospora	1,083	39.48
Paramecium – Dictyostelium	1,477	38.94
Paramecium – Thalassiosira	1,164	40.81
Paramecium – Cyanidioschyzon	880	40.48
Paramecium – Tetrahymena	**4,225**	**41.06**
Paramecium – Saccharomyces	995	39.92

Table S7. Characterization of paralogous genes for each WGD.

Duplication events	Number of copies	Number of genes	Average %ID (at amino acid level)	Average %ID (nucleotide level)	Average Ka	Average Ks	Average Ka/Ks	Average Ks/Ka	Average number of cDNAs
Recent WGD		35,503	82.8	82.0	0.10	2.17	0.08	29.88	2.1
	1 copy	11,451	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.2
	2 copies	24,052	82.77	81.98	0.10	2.17	0.08	29.88	2.5
Intermediary WGD		31,129	66.4	67.3	0.26	>5	n.c.	n.c.	2.1
	1 copy	6,625	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.2
	2 copies	**13,422**	55.3	61.0	0.36	>5	n.c.	n.c.	1.8
	3 copies	5,682	61.7	64.5	0.32	>5	n.c.	n.c.	1.6
	4 copies	5,400	**71.3**	**70.2**	**0.20**	>5	n.c.	n.c.	**4.3**
							n.c.	n.c.	
Old WGD		18.792	49.5	55.9	0.51	>5	n.c.	n.c.	2.2
	1 copy	3,582	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1.3
	2 copies	**6,964**	37.5	49.3	0.61	>5	n.c.	n.c.	1.6
	3 copies	3,084	41.2	51.2	0.62	>5	n.c.	n.c.	1.8
	4 copies	3,204	42.7	51.5	0.61	>5	n.c.	n.c.	3.8
	5 copies	840	45.6	53.6	0.52	>5	n.c.	n.c.	2.3
	6 copies	714	50.2	56.1	0.52	>5	n.c.	n.c.	3.7
	7 copies	196	51.6	56.7	0.49	>5	n.c.	n.c.	4.0
	8 copies	208	**76.0**	**72.6**	**0.22**	>5	n.c.	n.c.	**15.3**

Table S8. Distribution of DNA sequence divergence between paralogous sequences, for pairs of paralogs where one copy has been pseudogenized and for pairs of paralogs where both genes are functional.

Identity percent	Paralogous – Pseudogene (number of cases)	Paralogous – Paralogous
50-59	88.81 (1119)	1.27
60-69	9.60 (121)	3.67
70-79	1.03 (13)	19.59
80-89	0.48 (6)	59.23
90-100	0.08 (1)	16.24

Table S9. Stoichiometry in MIPS complexes. P-values were estimated by randomizing complex composition 1 000 000 times and counting the frequency of cases with a higher percentage of conserved stoichiometry than in the data.

Duplication events	Complexes size	Number of complexes with at least 2 Paramecium genes	Number of complexes with conserved stoichiometry	p-values
Recent WGD		599	265 (44%)	2.6×10^{-2}
	2	196		
	3	104		
	4	80		
	5-10	147		
	11+	72		
Intermediary WGD		562	114 (20%)	1.5×10^{-3}
	2	174		
	3	113		
	4	83		
	5-10	134		
	11+	58		
Old WGD		443	106 (24%)	1.2×10^{-5}
	2	166		
	3	96		
	4	56		
	5-10	104		
	11+	21		

Table S10. Stoichiometry in complexes from[31]. P-values were estimated by randomizing complex composition 1 000 000 times and counting the frequency of cases with a higher percentage of conserved stoichiometry than in the data.

Duplication events	Complexes size	Number of complexes with at least 2 Paramecium genes	% of complexes with conserved stoichiometry	p-values
Recent WGD		109		
	2	66		
	3	22	74 (68%)	4.3×10^{-4}
	4	9		
	5-9	12		
Intermediary WGD		101		
	2	67		
	3	15	43 (43%)	2.4×10^{-4}
	4	9		
	5-9	10		
Old WGD		61		
	2	44		
	3	9	26 (43%)	2.5×10^{-3}
	4	7		
	5	1		

Table S11. Conservation of stoichiometry in a complex. Nine *Paramecium* orthologs of the rRNA maturation complex (440.30.20) from yeast are shown. Eight have 2 duplicated genes, but three of them (bold) have retained paralogs from the intermediary duplication and lost both paralogs from the recent duplication.

Yeast protein	description in yeast	Copy number in Paramecium	paralogs retained from the recent WGD	Paralogs retained from the intermediary WGD	Paralogs retained from the old WGD
YDL014W	Fibrillarin	4	2	2	0
YER082C	U3 snoRNP Protein	2	2	0	0
YGR159C	Nuclear localization sequence binding protein	2	2	0	0
YHR089C	Nucleolar rRNA Processing Protein	2	2	0	0
YLL008W	RNA Helicase (DEAD box family)	2	**0**	**2**	0
YLR222C	U3snoRNP Protein	2	2	0	0
YLR223C	Pre-rRNA processing machinery control protein	2	**0**	**2**	0
YNR052C	Glucose repression required Protein	2	**0**	**2**	0
YOR048C	5’-3’ exoribonuclease	2	2	0	0

Table S12. Proportion of duplicated genes in some metabolic pathways.

Pathway Name	Pathway Coverage (%)	Number of EC numbers found	Number of genes	Proportion of duplicated genes (%)		
				Recent WGD	Intermediary WGD	Old WGD
Purine metabolism	41	41	297	91.8367	21.7391	NF
Pyruvate metabolism	50	33	122	78.5714	23.913	5.08475
Glycine, serine and threonine metabolism	46.6667	28	100	84	9.375	NF
Pyrimidine metabolism	45.1613	28	152	87.931	38.4615	NF
Glycolysis / Gluconeogenesis	65	26	139	93.9394	NF	NF
Butanoate metabolism	46.1538	24	138	80.9524	NF	4.08163
Starch and sucrose metabolism	32.4324	24	54	80	13.3333	NF
Glutamate metabolism	63.8889	23	83	73.0769	5	NF
Glycerophospholipid metabolism	33.8235	23	64	60.7143	10	NF
Aminoacyl-tRNA synthetases	100	21	60	93.3333	34.4828	NF
Valine, leucine and isoleucine degradation	63.6364	21	106	96	8.57143	5.55556
Arginine and proline metabolism	27.3973	20	78	96.4286	7.40741	NF
Nitrogen metabolism	31.746	20	75	77.7778	31.5789	7.69231
Tyrosine metabolism	27.7778	20	72	91.6667	12.1212	NF
Alanine and aspartate metabolism	50	19	59	69.2308	35	18.4211
Propanoate metabolism	42.2222	19	108	100	22.9167	NF
Fatty acid metabolism	64.2857	18	146	NC	11.1111	NF
Glyoxylate and dicarboxylate metabolism	31.0345	18	64	92.3077	32.3529	NF
Fructose and mannose metabolism	27.4194	17	67	90.4762	NF	NF
Phenylalanine metabolism	37.7778	17	76	96.4286	12.9032	8
Phenylalanine, tyrosine and tryptophan biosynthesis	54.8387	17	43	84.6154	36.8421	NF
Citrate cycle (TCA cycle)	69.5652	16	102	96.1538	47.5	NF
Lysine degradation	28.3019	15	221	66.6667	3.57143	NF
Pentose phosphate pathway	44.1176	15	60	92.3077	11.1111	NF
Phosphatidylinositol signaling system	65.2174	15	118	100	15.7895	NF

NF, not found. No duplicated genes have been found in the concerned pathway.
NC, not computed. No genes have been found in the concerned pathway.

Table S13. Annotation of the genes with low Ks values between old duplicates. RP = Ribosomal Protein.

Paralogous genes family	Annotation	Ks*	Aminoacids identity*	Expression level**	copy number
All genes	n.d.	>5 (saturated)	49.5%	2.2	2.04
1	RP S9	1.25	97.2%	12	6
2	RP L37	1.35	91.3%	7.7	7
3	RP S8	1.8	87.4%	8.9	7
4	RP S23	0.5	95.6%	6.4	8
5	RP L18	1.3	98.9%	8.5	8
6	RP L35	1.19	86.9%	11.8	8
7	RP L13	1.27	88.4%	10.4	8
8	RP L31	2.39	93.1%	8.3	8
9	RP L36	1.55	100%	2.7	6
10	RP L8	2.33	94.2%	21.3	6
11	RP L17	1.31	93.5%	10.4	8
12	RP L35	1.4	91.2%	9	8
13	RP S7	1.91	81.4%	18.4	8
14	RP L12	0.85	100%	8.9	6
15	RP L27	1.02	96.6%	9.7	8
16	RP L21	1.78	96.8%	13.2	5
17	RP S11	1.05	96.8%	10.8	8
18	RP L2	1.4	94.7%	33.3	7
19	RP S13	0.98	100%	8.9	8
20	RP S15	1.52	100%	9.7	6
21	RP L15	1.1	99.5%	11.8	6
22	RP L9	2.75	76.5%	16.2	6
23	RP S12	1.21	87.6%	12.7	7
24	RP S19	1.01	94.1%	7.4	8
25	Histone H2A	1.18	95.3%	4.5	4
26	Histone H3	0.51	100%	4.2	5
27	Glutathione S Transferase	2.4	68.7%	33	3
28	centrin 1A	1.89	96.7%	5.4	5
29	Beta tubulin	0.49	100%	161	3
30	Succinate Dehydrogenase alpha subunit	2.97	96.7%	6.8	6
31	translation elongation factor 2	2.68	80.2%	40.8	5
32	cAMP dependent kinase	2.61	96.6%	2.8	8
33	Ammonium Transporter	2.13	68.1%	19.3	7
34	Trichocyst Matrix Protein	0.38	97.4%	86	6
35	Trichocyst Matrix Protein	0.72	88.4%	81.3	8
36	Trichocyst Matrix Protein	0.66	93%	114.6	5
37	Trichocyst Matrix Protein	1.31	89.6%	29.8	6
38	Trichocyst Matrix Protein	0.7	96%	72.1	8
39	Homologous to Tetrahymena granule lattice protein 1	1.13	90.8%	66.3	8
40	Homologous to Tetrahymena granule lattice protein 4	2.1	81.1%	157.8	4
41	unknown	2.33	92.3%	2.3	3
42	unknown	1.89	95.1%	9.8	6
43	unknown	1.7	94.8%	16.4	5
44	unknown	1.79	96.5%	6.2	6
45	unknown	1.53	89.9%	22.6	8
46	unknown	1.9	90.9%	39.3	6
47	unknown	2.52	65.9%	3.8	6
48	unknown	2.3	77.2%	4.7	6
49	unknown	2.81	97.7%	6.5	6

* The lowest Ks value is shown. The a.a. identity also refered to that of the pair with the lowest Ks value.

** Expression levels are averaged across all duplicates.

Table S14. Frequency of asymmetric pairs according to the substitution rate of the slowly-evolving copy (copy1).

		# Pairs of duplicates	Average rate of copy1 (± standard deviation)	% pairs with asymmetric rate
Recent WGD	All genes	2297	0.07 ± 0.07	15.9%
	Slowly evolving genes	1727	0.04 ± 0.03	15.6%
	Intermediate genes	447	0.14 ± 0.03	16.8%
	Fast evolving genes	123	0.29 ± 0.09	17.1%
Intermediary WGD	All genes	293	0.18 ± 0.13	21.2%
	Slowly evolving genes	84	0.05 ± 0.03	17.9%
	Intermediate genes	100	0.15 ± 0.03	22.0%
	Fast evolving genes	109	0.31 ± 0.10	22.9%

Table S15. Average rate of evolution of the slowly-evolving copy (copy1) and fast evolving copy (copy2) for pairs of duplicates showing asymmetric or non-asymmetric rates of evolution.

	Pattern of evolution	# Pairs of duplicates	Average rate of copy1 (± standard deviation)	Average rate of copy2 (± standard deviation)
Recent WGD	No asymmetry	1931	0.07 ± 0.07	0.10 ± 0.09
	Asymmetry	366	0.07 ± 0.08	0.20 ± 0.20
Intermediary WGD	No asymmetry	231	0.18 ± 0.13	0.22 ± 0.14
	Asymmetry	62	0.19 ± 0.13	0.40 ± 0.24