Macronuclear Genome Sequence of the Ciliate *Tetrahymena thermophila*, a Model Eukaryote

Jonathan A. Eisen[1¤a*], Robert S. Coyne[1], Martin Wu[1], Dongying Wu[1], Mathangi Thiagarajan[1], Jennifer R. Wortman[1], Jonathan H. Badger[1], Qinghu Ren[1], Paolo Amedeo[1], Kristie M. Jones[1], Luke J. Tallon[1], Arthur L. Delcher[1¤b], Steven L. Salzberg[1¤b], Joana C. Silva[1], Brian J. Haas[1], William H. Majoros[1¤c], Maryam Farzad[1¤d], Jane M. Carlton[1¤e], Roger K. Smith Jr.[1¤f], Jyoti Garg[2], Ronald E. Pearlman[2,3], Kathleen M. Karrer[4], Lei Sun[4], Gerard Manning[5], Nels C. Elde[6¤g], Aaron P. Turkewitz[6], David J. Asai[7], David E. Wilkes[7], Yufeng Wang[8], Hong Cai[9], Kathleen Collins[10], B. Andrew Stewart[10], Suzanne R. Lee[10], Katarzyna Wilamowska[11], Zasha Weinberg[11¤h], Walter L. Ruzzo[11], Dorota Wloga[12], Jacek Gaertig[12], Joseph Frankel[13], Che-Chia Tsao[14], Martin A. Gorovsky[14], Patrick J. Keeling[15], Ross F. Waller[15¤j], Nicola J. Patron[15¤j], J. Michael Cherry[16], Nicholas A. Stover[16], Cynthia J. Krieger[16], Christina del Toro[17¤k], Hilary F. Ryder[17¤l], Sondra C. Williamson[17], Rebecca A. Barbeau[17¤m], Eileen P. Hamilton[17], Eduardo Orias[17]

1 The Institute for Genomic Research, Rockville, Maryland, United States of America, **2** Department of Biology, York University, Toronto, Ontario, Canada, **3** Centre for Research in Mass Spectrometry, York University, Toronto, Ontario, Canada, **4** Department of Biological Sciences, Marquette University, Milwaukee, Wisconsin, United States of America, **5** Razavi-Newman Center for Bioinformatics, The Salk Institute for Biological Studies, San Diego, California, United States of America, **6** Department of Molecular Genetics and Cell Biology, University of Chicago, Chicago, Illinois, United States of America, **7** Department of Biology, Harvey Mudd College, Claremont, California, United States of America, **8** Department of Biology, University of Texas at San Antonio, San Antonio, Texas, United States of America, **9** Department of Electrical Engineering, University of Texas at San Antonio, San Antonio, Texas, United States of America, **10** Department of Molecular and Cellular Biology, University of California Berkeley, Berkeley, California, United States of America, **11** Department of Computer Science and Engineering, University of Washington, Seattle, Washington, United States of America, **12** Department of Cellular Biology, University of Georgia, Athens, Georgia, United States of America, **13** Department of Biological Sciences, University of Iowa, Iowa City, Iowa, United States of America, **14** Department of Biology, University of Rochester, Rochester, New York, United States of America, **15** Canadian Institute for Advanced Research, Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada, **16** Department of Genetics, Stanford University, Stanford, California, United States of America, **17** Department of Molecular, Cellular, and Developmental Biology, University of California Santa Barbara, Santa Barbara, California, United States of America

The ciliate *Tetrahymena thermophila* is a model organism for molecular and cellular biology. Like other ciliates, this species has separate germline and soma functions that are embodied by distinct nuclei within a single cell. The germline-like micronucleus (MIC) has its genome held in reserve for sexual reproduction. The soma-like macronucleus (MAC), which possesses a genome processed from that of the MIC, is the center of gene expression and does not directly contribute DNA to sexual progeny. We report here the shotgun sequencing, assembly, and analysis of the MAC genome of *T. thermophila*, which is approximately 104 Mb in length and composed of approximately 225 chromosomes. Overall, the gene set is robust, with more than 27,000 predicted protein-coding genes, 15,000 of which have strong matches to genes in other organisms. The functional diversity encoded by these genes is substantial and reflects the complexity of processes required for a free-living, predatory, single-celled organism. This is highlighted by the abundance of lineage-specific duplications of genes with predicted roles in sensing and responding to environmental conditions (e.g., kinases), using diverse resources (e.g., proteases and transporters), and generating structural complexity (e.g., kinesins and dyneins). In contrast to the other lineages of alveolates (apicomplexans and dinoflagellates), no compelling evidence could be found for plastid-derived genes in the genome. UGA, the only *T. thermophila* stop codon, is used in some genes to encode selenocysteine, thus making this organism the first known with the potential to translate all 64 codons in nuclear genes into amino acids. We present genomic evidence supporting the hypothesis that the excision of DNA from the MIC to generate the MAC specifically targets foreign DNA as a form of genome self-defense. The combination of the genome sequence, the functional diversity encoded therein, and the presence of some pathways missing from other model organisms makes *T. thermophila* an ideal model for functional genomic studies to address biological, biomedical, and biotechnological questions of fundamental importance.

Citation: Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, et al. (2006) Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol 4(9): e286. DOI: 10.1371/journal.pbio.0040286

Introduction

Tetrahymena thermophila is a single-celled model organism for unicellular eukaryotic biology [1]. Studies of *T. thermophila* (referred to as *T. pyriformis* variety 1 or syngen 1 prior to 1976 [2]) have contributed to fundamental biological discoveries such as catalytic RNA [3], telomeric repeats [4,5], telomerase [6], and the function of histone acetylation [7]. *T. thermophila* is advantageous as a model eukaryotic system because it grows rapidly to high density in a variety of media and conditions, its life cycle allows the use of conventional tools of genetic analysis, and molecular genetic tools for sequence-enabled experimental analysis of gene function have been developed [8,9]. In addition, although it is unicellular, it possesses many core processes conserved across a wide diversity of eukaryotes (including humans) that are not found in other single-celled model systems (e.g., the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*).

T. thermophila is a member of the phylum Ciliophora, which also includes the genera *Paramecium*, *Oxytricha*, and *Ichthyoph-*

thirius. A cartoon showing the phylogenetic position of *T. thermophila* relative to other eukaryotes for which the genomes have been sequenced is shown in Figure 1. The ciliates are one of three major evolutionary lineages that make up the alveolates. The other two lineages are dinoflagellates and the exclusively parasitic apicomplexa, which includes the *Plasmodium* species that cause malaria. Although experimental tools are improving for the apicomplexa [10–12], they can still be challenging to work with, and in some situations *T. thermophila* can serve as a useful "distant cousin" model for this group [13].

As is typical of ciliates, *T. thermophila* cells exhibit nuclear dimorphism [14]. Each cell has two nuclei, the micronucleus (MIC) and the macronucleus (MAC), containing distinct but closely related genomes. The MIC is diploid and contains five pairs of chromosomes. It is the germline, the store of genetic information for the progeny produced by conjugation in the sexual stage of the *T. thermophila* life cycle. Conjugation involves meiosis, fusion of haploid MIC gametes to produce a new zygotic MIC, and differentiation of new MACs from mitotic copies of the zygotic MIC (for details, see [15]). After formation of the MAC, cells reproduce asexually until the next sexual conjugation. During this asexual growth, all gene expression occurs in the MAC, which is thus considered the somatic nucleus.

The MAC genome derives from that of the MIC, but the two genomes are quite distinct. During MAC differentiation, several types of developmentally programmed DNA rearrangements occur [16,17] (Figure 2). One such rearrangement is the deletion of segments of the MIC genome known as internally eliminated sequences (IESs). It is estimated that approximately 6,000 IESs are removed, resulting in the MAC genome being an estimated 10% to 20% smaller than that of the MIC [18]. A key aspect of the process is the preferential removal of repetitive DNA, which results in 90% to 100% of MIC repeats being eliminated [19,20]. Thus the process can be considered analogous to and more extreme than other forms of repeat element silencing phenomena such as repeat-induced point mutation (RIP) in *Neurospora* and heterochromatin formation [21,22]. A second programmed DNA rearrangement is the site-specific fragmentation at each location of the 15–base pair (bp) chromosome breakage sequence (Cbs) [23–25]. During fragmentation, sections of the MIC genome containing each Cbs, as well as up to 30 bp on either side, are deleted [26]. Telomeres are then added to each new end [27], generating some 250 to 300 MAC chromosomes [28,29].

Another process that occurs during MAC differentiation is the amplification of the number of copies of the MAC chromosomes. The rDNA chromosome, which encodes the 5.8S, 17S, and 26S rRNAs, is maintained at an average of 9,000 copies per MAC [30]. Six other chromosomes that have been examined are each maintained at an average of 45 copies per MAC [31]. During asexual reproduction, the MAC divides amitotically, with apparently random distribution of chromosome copies that behave as if acentromeric. In contrast, MIC chromosomes are metacentric [32] and are distributed mitotically [33,34]. Parental MAC DNA is not transmitted to sexual progeny, although it does have an epigenetic influence on postzygotic MAC genome rearrangement, mediated by RNA interference [35].

The *Tetrahymena* research community has coordinated an effort to develop genomic tools for *T. thermophila* [9,36]. The MAC genome was selected for initial sequencing because it contains all the expressed genes and because the complexity of the assembly process was expected to be reduced due to the lower amounts of repetitive DNA. These advantages, however, are countered by some complexities not seen in other eukaryotic genome projects, including the presence of several hundred medium-sized to small chromosomes, the possibility of unequal copy number of at least some chromosomes, the existence of polymorphisms that are generated during MAC development, and the inability to completely separate the MIC from the MAC prior to DNA isolation.

We report here on the shotgun sequencing, assembly, and analysis of the MAC genome of *T. thermophila* strain SB210, an inbred strain B derivative that has been extensively used for genetic mapping and for the isolation of mutants. We discuss how the complexities of sequencing the MAC were successfully addressed, as well as the biological and evolutionary implications of our analysis of the genome sequence.

Academic Editor: Mikhail Gelfand, Institute for Information Transmission Problems, Russian Federation

Received January 4, 2006; **Accepted** June 23, 2006; **Published** August 29, 2006

DOI: 10.1371/journal.pbio.0040286

Abbreviations: bp, base pairs; Cbs, chromosome breakage sequence; CM, covariance model; EST, expressed sequence tag; IES, internal eliminated sequence; ITR, inverted terminal repeat; MAC, macronucleus/macronuclear; MIC, micronucleus/micronuclear; ncRNA, noncoding RNA; RIP, repeat induced point mutation; SCI, single-cell isolation; Sec, selenocysteine; TE, transposable element; TGD, *Tetrahymena* Genome Database; TIGR, The Institute for Genomic Research; VIC, voltage-gated ion channel

* To whom correspondence should be addressed. E-mail: jaeisen@ucdavis.edu

¤a Current address: University of California Davis Genome Center, Section of Evolution and Ecology, School of Biological Sciences and Department of Medical Microbiology and Immunology, School of Medicine, University of California Davis, Davis, California, United States of America

¤b Current address: Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, United States of America

¤c Current address: Duke Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina, United States of America

¤d Current address: Agilent Technologies, Inc., Santa Clara, California, United States of America

¤e Current address: Department of Medical Parasitology, New York University School of Medicine, New York, New York, United States of America

¤f Current address: Dupont Agriculture and Nutrition, Wilmington, Delaware, United States of America

¤g Current address: Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America

¤h Current address: Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut, United States of America

¤j Current address: School of Botany, The University of Melbourne, Melbourne, Australia

¤k Current address: Meharry Medical College, Nashville, Tennessee, United States of America

¤l Current address: Dartmouth-Hitchcock Medical Center, Lebanon, New Hampshire, United States of America

¤m Current address: Lung Biology Center, University of California San Francisco, San Francisco, California, United States of America

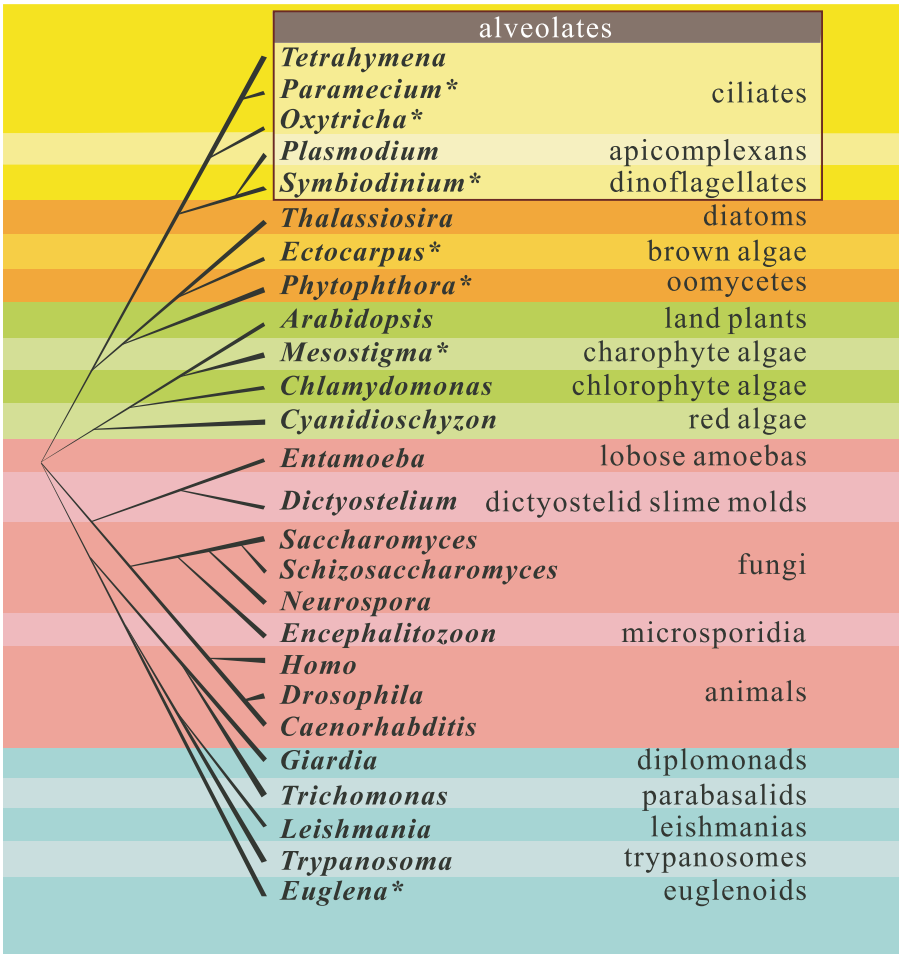

Figure 1. Unrooted Consensus Phylogeny of Major Eukaryotic Lineages

Representative genera are shown for which whole genome sequence data are either in progress (marked with asterisks *) or available. The ciliates, dinoflagellates, and apicomplexans constitute the alveolates (lighter yellow box). Branch lengths do not correspond to phylogenetic distances. Adapted from the more detailed consensus in [197].

DOI: 10.1371/journal.pbio.0040286.g001

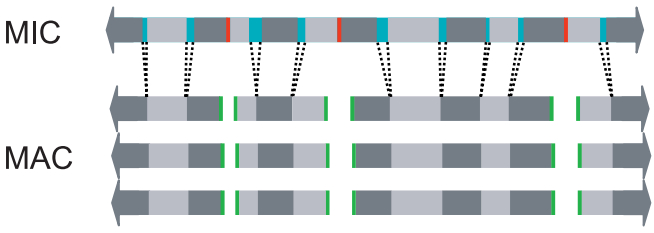

Figure 2. Relationship between MIC and MAC Chromosomes

The top horizontal bar shows a small portion of one of the five pairs of MIC chromosomes. MAC-destined sequences are shown in alternating shades of gray. MIC-specific IESs (internally eliminated sequences) are shown as blue rectangles, and sites of the 15-bp Cbs are shown as red bars (not to scale). Below the top bar are shown macronuclear chromosomes derived from the above region of the MIC by deletion of IESs, site-specific cleavage at Cbs sites, and amplification. Telomeres are added to the newly generated ends (green bars). Most of the MAC chromosomes are amplified to approximately 45 copies (only three shown). Through the process of phenotypic assortment, initially heterozygous loci generally become homozygous in each lineage within approximately 100 vegetative fissions. Polymorphisms located on the same MAC chromosome tend to co-assort.

DOI: 10.1371/journal.pbio.0040286.g002

Results/Discussion

Genome Assembly and General Chromosome Structure

Sequencing and assembly. Using physical isolation methods, MAC were purified from a culture of *T. thermophila* strain SB210 and used to create multiple differentially sized shotgun sequencing libraries (Table S1). Construction of large (greater than 10 kb) insert libraries was not successful—a common problem in working with AT-rich genomes. Approximately 1.2 million paired end sequences were generated from the libraries and assembled using the Celera Assembler [37]. In an initial assembly, the mitochondrial genome (mtDNA; which was present due to some contamination of the MAC preparation with mitochondria) and the highly amplified rDNA chromosome did not assemble well compared to the published sequences of these molecules [38,39]. This was probably because contigs from these molecules had higher depths of coverage than those from other chromosomes, which caused the Celera Assembler to treat them as repetitive DNA. Thus we divided sequence reads into three bins (mtDNA, rDNA, and bulk MAC DNA) and generated assemblies for each bin separately. This resulted in a moderate improvement, and the three separate assemblies

were thus used for all subsequent analyses. Detailed sequence and assembly information is presented in Tables 1 and S2.

The bulk MAC assembly contains 1,971 scaffolds (contigs that have been linked into larger pieces by mate pair information) with a total estimated span of 104.1 Mb. Perhaps most important, using a combination of computational and experimental identification of telomeres, we have found that many scaffold ends correspond to chromosome ends. One hundred twenty-five scaffolds, encompassing 44% of the assembled genome length, are telomere-capped at both ends and thus likely represent complete MAC chromosomes. One hundred twenty additional scaffolds, encompassing another 31% of the genome, are telomere-capped at one end (Tables 1 and S3).

Assembly accuracy and completeness. Overall, all analyses indicate that the bulk MAC assemblies are highly accurate. For example, all 75 MAC loci that are in distinct genetic coassortment groups (and thus should be on different chromosomes [40]) map to different scaffolds, and all pairs of loci that coassort (and thus should be on the same chromosome) either map to the same scaffold or to two non–fully capped scaffolds whose cumulative size is less than that of the corresponding MAC chromosome (Table S4). For the 24 completely assembled chromosome scaffolds for which we know the corresponding chromosome physical size, there is a very strong correlation between physical size and assembly length. In addition, there are no cases where a scaffold is significantly longer than the physical size of the corresponding chromosome (Figure 3A). Finally, all of the 96 MIC sequences known to be adjacent to Cbs sites [24,41,42] that matched to a MAC scaffold did so only at the scaffold's end.

The general accuracy of the assemblies indicates that many of the potential difficulties discussed in the Introduction were not significant. For example, we see little evidence for polymorphism among reads, which is likely a reflection of the use of an inbred strain and the process of phenotypic assortment, which leads to whole-genome MAC homozygous lineages [43]. Also, searches for known MIC-specific sequences indicate that the amount of MIC contamination is very low (e.g., Cbs junctions are at 0.044× coverage which is approximately 200-fold less than the bulk MAC chromosomes) and limited to small contigs (most less than 5 kb). The uniform depth of contig coverage and accuracy of assemblies also suggest that the chromosomes are present in roughly similar copy number and that only limited amounts of repetitive DNA are present in the MAC, both of which are discussed further below.

The total scaffold length is much smaller than the predicted genome size of 180 to 200 Mb [14]. Given the accuracy of the assemblies, the large number of chromosomes partially or completely capped, and the fact that all (more than 200) known MAC DNA sequences are found in the assemblies, we conclude that the assemblies represent a very large (more than 95%) fraction of the genome. We conclude therefore that previous genome size estimates were inaccurate (which is not surprising given that they were made almost 30 years ago) and that the genome is close to 105 Mb in size. It is possible, however, that some chromosomes or regions were underrepresented in our libraries due to purification or cloning bias, and thus one cannot infer the absence of any particular gene or feature simply due to its absence from our current assemblies.

Table 1. Important Genome Statistics

Category	Number
Sequence reads	
Total	1,180,981
Reads in contigs	1,137,759 (96.3% of total)
Estimated coverage	9.08-fold
Contigs	
In scaffolds	2,955
Total bp in contigs	103,927,049 bp
Total bp in contigs >10 kb	99,668,989 bp (95.9% of total)
Maximum contig size[a]	715,652 bp
Scaffolds	
Total	1,971
Total bases in scaffolds	103,927,049 bp
Span of scaffolds	104,194,423 bp
Longest scaffold[a]	2,214,258 bp
Average GC content	22%
Telomere reads and scaffolds	
Telomere-containing reads[b]	4,058
Telomere reads linked to scaffold ends	3,328 (82% of total)
Telomere-capped scaffold ends	370 (82% of total)[d]
Telomere coverage[c]	8.99-fold
Scaffolds capped at both ends	125
Base pairs in two-cap scaffolds	45,191,229 (44% of total)
Scaffolds capped at only one end	120
Base pairs in one-cap scaffolds	31,827,449 (31% of total)

[a]Potentially limited by natural fragmentation of the MAC genome.
[b]Non-rDNA chromosomes.
[c]For telomere-capped ends.
[d]Assuming a total of 450 ends (225 MAC chromosomes).
DOI: 10.1371/journal.pbio.0040286.t001

Estimating the number of MAC chromosomes. The total number of MAC chromosomes is unknown. The telomere-capping of scaffolds allows us to place a minimum boundary on this number at 185 (125 plus half of 120). One way of estimating the actual number is through analysis of the non–rDNA telomere-containing reads; 3,328 such reads can be linked to a total of 370 scaffold ends. This corresponds to approximately 9-fold coverage (3,328/370), which is not significantly different from the bulk MAC chromosome coverage of 9.08, indicating that there is no significant underrepresentation of telomere reads (Tables 1 and S3). Thus since there are 4,058 such reads total (the others could not be linked), we estimate that there are approximately 451 telomere ends (4,058/9), and thus that there are approximately 225 chromosomes (451/2). An independent estimate of the actual chromosome number can be made by assuming that the size distribution of fully capped chromosomes (see Figure 3B) is representative of the genome as a whole. Since these 125 capped chromosomes represent 43.5% of the total assembly length, this would predict 287 chromosomes in total (125/0.435). This is likely to be an overestimate, since larger chromosomes are statistically less likely to be in the completely assembled set. Indeed, the average size of completely assembled chromosomes is 359 kb, whereas estimates of the average MAC chromosome size obtained through pulsed-field gel electrophoresis are substantially higher [29,41]. Thus, we conclude that there are between 185 and 287 chromosomes, most likely somewhere near 225.

Absence of many standard global features of eukaryotic chromosomes. We note that we searched for but could not

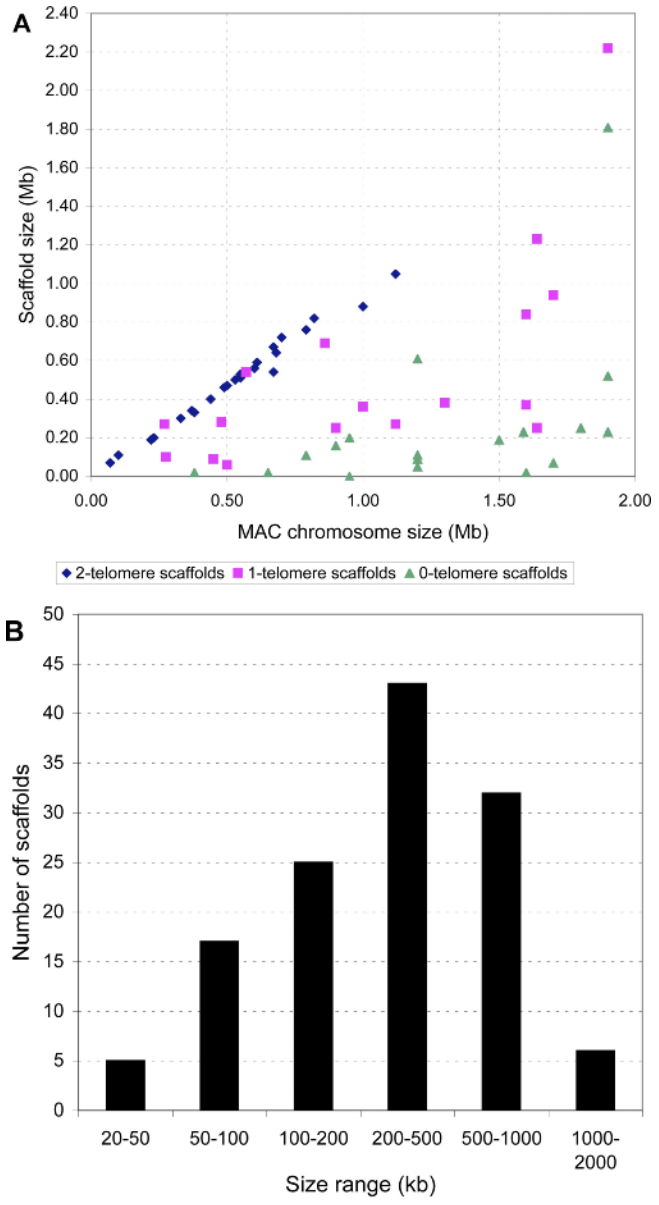

Figure 3. Scaffold Sizes

(A) Scaffold sizes versus MAC chromosome size. Blue diamonds represent scaffolds capped by telomeres on both ends. Red squares and green triangles represent incomplete scaffolds capped by telomeres at one or neither end, respectively.

(B) Size distribution of scaffolds capped by telomeres on both ends.

DOI: 10.1371/journal.pbio.0040286.g003

find many of what are considered standard global features of eukaryotic chromosomes. For example, we could not find sequence or structural features shared across multiple chromosomes that could be considered candidates for centromeric regions. This is consistent with experimental studies [44]. In addition, although in many eukaryotes certain genes and repeat elements cluster near telomeres [45–51], we cannot detect any such clustering here. This is not because there is no variation in these features; for example, GC content (Figure S1) and gene density (Figure S2) do vary greatly. Instead, the absence of similar global structure between MAC chromosomes is likely due to the absence of the processes that help generate the key features of normal eukaryotic chromosomes (e.g., mitosis and meiosis, which in *T. thermophila* are confined to the MIC).

MAC chromosome copy number is uniform. The high quality and completeness of the assemblies suggest that copy number variation among at least most MAC chromosomes is relatively small since otherwise the assembler would have treated contigs from overrepresented chromosomes as repetitive DNA. Such uniform copy number is consistent with genetic experimental data for six chromosomes [31], but its generality for all chromosomes has been unknown. We realized that the relative chromosome copy number could be estimated from depth of coverage in our assemblies (assuming that cloning and sequencing success were relatively random). When all scaffolds are examined, the depth of coverage is remarkably uniform (Figure 4). The decrease in uniformity and coverage seen as scaffold size decreases is likely a reflection of both chance low coverage of some regions and some of the small scaffolds being MIC contaminants. When only scaffolds capped by telomeres at both ends are included in the analysis, observed sequence coverage is even more uniform (red diamonds in Figure 4). Although we cannot rule out that some smaller, incompletely assembled chromosomes are maintained at different copy numbers, the observed uniformity indicates that the replication and/or segregation of most or all bulk MAC chromosomes is under coordinated regulation.

General Features of Predicted Protein Coding Genes and Noncoding RNAs

Protein coding gene predictions. We identified 27,424 putative protein-coding genes in the genome (Table 2), a high number for a single-celled species. These gene models were tested by aligning expressed sequence tags (ESTs) to the genome assemblies using PASA [52]. We note that most of these ESTs were generated after the models were built (Table S5). Of the 9,122 EST clusters identified, most have either no conflicts with the gene models (49.5%) or relatively small ones (17.7% have a missed exon and 9.8% suggest the models need to be merged or split). Only 408 (4.4%) clusters are intergenic relative to the gene models. Although these could represent

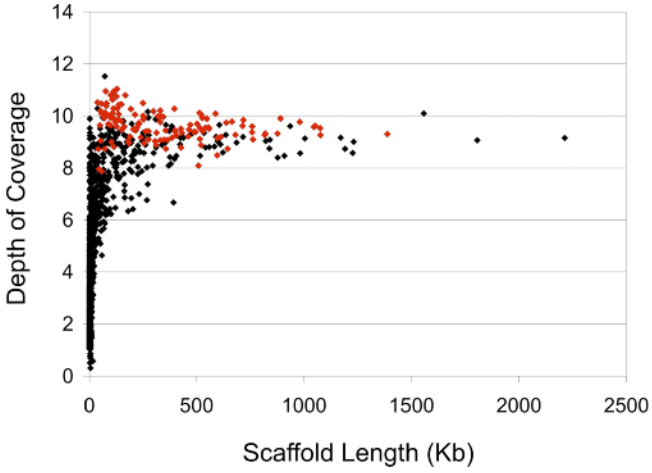

Figure 4. Depth of Coverage versus Scaffold Size

Black diamonds indicate all scaffolds; red diamonds, scaffolds capped with telomeres on both ends.

DOI: 10.1371/journal.pbio.0040286.g004

Table 2. Characteristics of Ab Initio Predicted Genes

Feature	Average (bp)	Minimum (bp)	Maximum (bp)	%GC
Genes	1,815.4	27	47,334	22.3
Exons	420.6	3	14,390	27.6
Introns	165.2	26	3,116	16.3
Intergenic regions	1,422.5	22	17,406	17.8

DOI: 10.1371/journal.pbio.0040286.t002

missed genes or gene regions, they could also be noncoding RNAs (ncRNAs) or genomic DNA contamination of cDNA libraries. In addition, the predicted and EST-derived introns are quite similar in size distribution except at the short and long extremes (Figure S3), GC content (16.3% versus 16.7%), and splice sites [only a small number (85) of EST-based introns have exceptions to the 5′-GT...AG-3′ junctions assumed by the model—these could simply be sequencing errors]. These analyses indicate that the gene models are relatively robust and should be more than sufficient for making general predictions about the coding potential of this species.

Two other lines of evidence suggest the predicted gene number is not inflated. First, a large number of the predicted genes have matches to known or predicted genes from other species (14,916 have a BLASTP match with an E-value better than 10^{-10}), and second, experimental studies of mRNA complexity predict transcription of at least 25,000 genes of an average size of 1,200 bp [53]. We also note that the sequence of the largest MAC chromosome of another ciliate, *Paramecium tetraurelia,* indicates a high coding density, and extrapolation to the complete genome predicts at least 30,000 protein-coding genes [54].

ncRNAs and the use of all 64 codons to code for amino acids. The ncRNAs found in the genome are listed in Table S6. We call attention to a few new findings. Of the 174 putative 5S rRNA genes (Table S6A), 19 do not correspond to any of the four previously reported *T. thermophila* sequences [55,56]. These 19 differ from one another by single nucleotide substitutions at 34 positions, as well as by various insertions, deletions, and truncations and may represent pseudogenes. In addition, there are two forms of U2 snRNA present (Table S6C), which we have termed U2 (four genes) and U2var (five genes). Functional RNA gene families are expressed ubiquitously during the *T. thermophila* life cycle and under stress conditions as well (representative data shown in Figure S4). The largest class is tRNAs with 700 identified (Tables S6B and S6D), a number consistent with hybridization-based estimates [57].

One of the more unusual features of *T. thermophila* and certain other ciliates is the use of an alternative genetic code in which the canonical stop codons UAG and UAA code for glutamine [58]. The importance and age of this alternative code are reflected in the genome by the presence of 39 tRNAs for these codons. Remarkably, analysis of the genome has also revealed the presence of a tRNA that is predicted to decode the remaining stop codon, UGA. Multiple lines of evidence indicate that this is a functioning tRNA for selenocysteine (Sec), the so-called 21st amino acid. In those eukaryotic species that use Sec, most UGA codons still cause translation termination while those mRNAs that encode Sec-containing peptides have a characteristic stem-loop sequence motif in the 3′ UTR region that directs Sec incorporation [59,60]. The putative *T. thermophila* tRNA-Sec was identified by analysis of the genome sequence and shown to be transcribed and acylated [61], and we have found that it is expressed and charged and that its charging may be under distinct regulatory control from other tRNAs (Figure S4A). In addition, we identified six *T. thermophila* genes with in-frame UGA codons that align (after editing of the gene models) with known Sec codons of their homologs from other eukaryotic species and that have the stem-loop consensus and thus are likely to encode selenoproteins. Thus we conclude that UGA is almost certainly translated into Sec, which would make *T. thermophila* the first organism known to use all 64 triplet codons to specify amino acid incorporation.

Genome Evolution

Codon and amino acid usage bias. Although *T. thermophila* can use all 64 codons, it does not use all equally. The most significant aspect of the codon usage in this species is that the AT-rich codons tend to be used more frequently than others [62,63]. Thus although the AT bias in the genome is strongest in noncoding regions, where selection is thought to be relaxed, it is seen even in coding regions. In fact, the AT pull is so strong in coding regions that amino-acid composition of proteins is shifted toward those coded by codons with high AT content, as seen in other species with extreme AT bias (e.g., [64]). Although the overall codon usage is biased against GC-rich codons, on a gene-by-gene level there is significant variation in the degree of bias. We have identified two dominant patterns to this gene-by-gene variation. The major pattern is that for most genes, the codons used are simply a reflection of the overall AT content of the gene (Figure 5). The variation among genes is due to genomewide variation in AT content (see Figure 5A), although we have been unable to discern a mechanism underlying this variation (e.g., there is no clustering of high or low AT genes near telomeres). There is, however, a less common pattern in the gene-by-gene variation that is very important. There exists a subset of genes (shown in red) that use a common preferred codon set that is different from that of the average gene, and the codons in this set are not strongly correlated to the genes' AT content. Although the existence of such a preferred codon set for this species has been reported [62,63], analysis of the genome allows the set and the genes that use it to be more precisely defined. In total, using a relatively conservative cutoff (Figure 5B), we have identified 232 such genes.

The use of preferred codons by a gene is thought to allow for more efficient or accurate translation [65]. This appears to be the case here as, of the predicted genes using the preferred subset, many have likely housekeeping functions, and, although they account for only 0.85% of all predicted genes, 12.5% of all ESTs map to them (Table S7). Although some do not have EST matches and theoretically could represent falsely predicted genes, it seems unlikely that spurious genes would use the preferred codon set. Thus we predict that these outlier genes are either highly expressed (in at least some of the conditions normally encountered by the organism) or have some critical function requiring accurate translation.

Codon usage differences between genes are thought to have only small fitness effects. For natural selection to

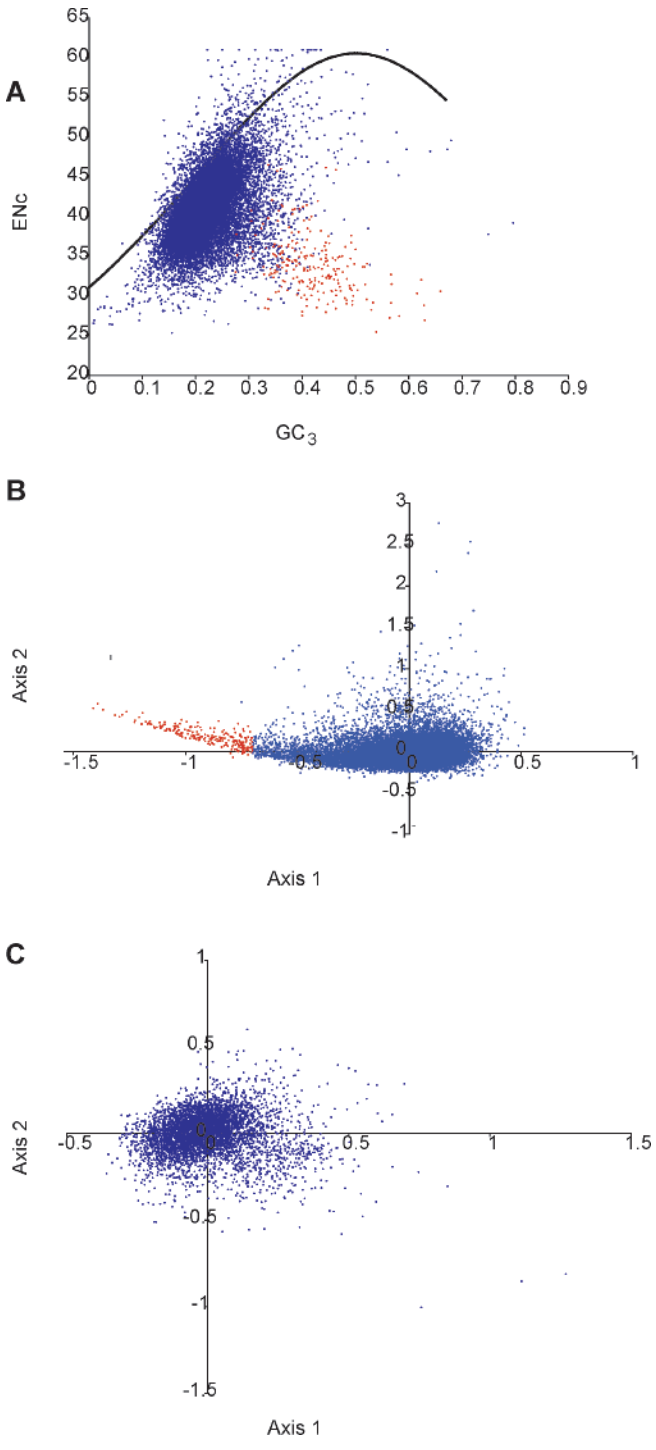

Figure 5. Codon Usage

(A) Effective number of codons (ENc; a measure of overall codon bias) for each predicted ORF is plotted versus GC3 (the fraction of codons that are synonymous at the third codon position that have either a guanine or a cytosine at that position). The upper limit of expected bias based on GC3 alone is represented by the black curve; most *T. thermophila* ORFs cluster below the curve [red dots as in (B)].

(B) Principal component analysis of relative synonymous codon usage in *T. thermophila*. The 232 genes in the tail of the comma-shaped distribution (those with the most biased codon usage) are colored red.

(C) Principal component analysis of relative synonymous codon usage in *P. falciparum*.

DOI: 10.1371/journal.pbio.0040286.g005

effectively work on codon usage differences and to thus create a preferred subset, factors that enhance genetic drift (e.g., small population sizes, inbreeding) must be weaker than the selective forces [66]. Thus although codon usage is probably under selective pressure in all species, not all are able to evolve preferred codon sets. For example, although it has a similar AT bias to *T. thermophila*, no preferred set could be detected in the apicomplexan *Plasmodium falciparum* (Figure 5C), possibly a reflection of its parasitic lifestyle and limited effective population size. The presence of a preferred subset in *T. thermophila* is likely a reflection of a large effective population size due to its free-living, sexually reproducing lifestyle (see [66,67] for additional discussion on the large population size of this species).

No plastid-derived genes can be identified. One question of particular interest that the *T. thermophila* genome might shed light on relates to the timing of the origin of the plastids found in apicomplexans and dinoflagellates, the other members of the alveolates [68,69]. Although the plastids in these lineages differ (e.g., that in apicomplexans, known as an apicoplast, is not even involved in photosynthesis), both are thought to be of red algal origin [70]. This has led to the proposal that the plastids in these lineages are the result of a single endosymbiotic event between an ancestor of apicomplexans and dinoflagellates and a red alga, with the algal nucleus being lost and the algal plastid being kept. A key question is whether this secondary endosymbiosis occurred before or after the ciliates split off from the other two lineages. The possibility that it occurred before the ciliate split is known as the chromalveolate hypothesis [71].

For the chromalveolate hypothesis to be correct, plastid loss would have to have occurred in ciliates, most likely at the base of the ciliate tree since no modern ciliates are known to harbor plastids. If the ancestor of ciliates once had a plastid, it is possible that some plastid-derived genes would have been transferred to the nuclear genome (as has occurred in many lineages including apicomplexans and dinoflagellates [72]), and furthermore that some such genes would still be found in *T. thermophila*. To test this possibility, we built phylogenetic trees of all genes in the genome and searched for those with a branching pattern consistent with plastid descent (see Materials and Methods). For *T. thermophila*, we do not see any signal for genes of plastid descent that rises above the noise seen in such automated phylogenetic analyses.

Several lines of evidence suggest that this is not a general flaw in the phylogenetic approach used here. For example, we have used the same approach to identify and catalog the plastid-derived genes in other lineages including the plant *Arabidopsis thaliana* and the apicomplexan *P. falciparum*. In addition, such an approach has been used to detect past endosymbioses in other eukaryotic lineages [73]. Finally, using the same approach we identified 91 likely mitochondrion-derived genes (Table S8) in the *T. thermophila* nuclear genome. This is significant because mitochondrion-derived genes are generally more difficult to identify than plastid-derived genes [74], in part because the plastid symbiosis was more recent [75].

Nevertheless, since it is possible that our phylogenomic screen might have missed some plastid-derived genes, we also did a targeted search for genes that might be expected to be retained, using the apicoplast as a model. Apicoplasts are involved in biosynthesis of fatty acids, isoprenoids, and heme.

Fatty acid and isoprenoid biosynthetic pathways are of special interest because the plastid-derived pathways are distinct from analogous pathways in the eukaryotic cytoplasm [76]. In the case of isoprenoid biosynthesis, genes for proteins in the canonical eukaryotic cytosolic mevalonate pathway are present as expected based on experimental studies [77–79], but no enzymes involved in the plastid-derived DOXP pathway were evident. For fatty acid biosynthesis, while *T. thermophila* does not require an exogenous supply of fatty acids for growth, no evidence for a complete version of a type I (normally cytosolic) pathway could be found. Although at least some genes for a type II pathway are present, these are insufficient for de novo fatty acid synthesis and appear more likely to be derived from the mitochondrion than a plastid.

Based on the general and targeted searches, we conclude that there is presently no evidence for a plastid or ancestrally plastid-derived genes in *T. thermophila*. This does not preclude the possibility that other ciliates have plastid-derived enzymes or even a plastid, but there is presently no evidence to suggest this despite extensive ultrastructural observations [80,81]. If ciliates do lack all evidence of a plastid, it could either mean that the hypothesized early origin of the chromalveolate plastid is incorrect or that an ancestor of *T. thermophila* (and perhaps all ciliates) lost its plastid and all detectable plastid-derived genes outright. The latter possibility is not without precedent, as some apicomplexans such as the Cryptosporidia have lost their apicoplasts and have few, if any, plastid-derived genes in their nuclear genomes [82,83]. This loss has been suggested to be the result of metabolic streamlining in response to its parasitic lifestyle. Resolving whether a plastid was present in the ancestor of ciliates will be important to our understanding of the evolution of plastids and their biochemical relationship with eukaryotic hosts.

IES excision targets foreign DNA rather than repetitive DNA *per se*. As discussed in the Introduction, there are multiple parallels between the IES excision process and other repeat element silencing phenomena such as RIP and heterochromatin formation. Despite these parallels, the processes differ significantly in their mechanisms of action and therefore likely have different short- and long-term evolutionary consequences. For example, in species with RIP, all repetitive DNA becomes a target for mutational inactivation, which has resulted in a drastic suppression of evolutionary diversification through gene duplication [84,85]. The IES excision process results in the exclusion of certain MIC DNA sequences from the transcriptionally active MAC. Experimental introduction of foreign transgenes into the MIC has shown that as MIC copy number increases, so does the efficiency of transgene excision [86]. One might therefore predict a similar suppression of gene duplication as in RIP. However, rather than targeting repetitive DNA per se, it has been proposed that IES excision specifically targets foreign DNA that has invaded the germline MIC but is not represented in the MAC [35,87,88]. MIC gene duplication and functional diversification should still be possible under this scenario as long as, at each conjugation event, the gene copies have not diverged in sequence enough to be recognized as foreign and excluded from the MAC; since sex is frequent in natural populations of *T. thermophila* [89], this should be the case. We therefore sought to use the genome sequence data to both test the foreign DNA hypothesis and to examine what the consequences of the IES excision process have been on the evolution of the *T. thermophila* genome.

Analysis of the genome reveals several lines of evidence that provide strong support for the foreign DNA hypothesis. First, small but nevertheless significant amounts of repetitive DNA are present in the MAC. This is best seen in analysis of the scaffolds that correspond to complete MAC chromosomes which are unlikely to contain MIC IES contamination. These scaffolds contain dispersed repeats that make up 2.3% of the total DNA. This means that some repetitive DNA bypasses the IES excision process. The second line of evidence comes from examining the small contigs and singletons (nonassembled sequences) in the assembly data. Known MIC-specific elements such as the REP and *Tlr*1 transposons [90,91] are found only in these small contigs, which are thus clearly enriched for MIC-specific DNA (and also for repetitive DNA; see Figure S5). In fact, the small contigs contain homologs of an unusually wide range of transposable element (TE) clades for a single-celled eukaryote [92,93] including many previously unreported in *Tetrahymena* (Table S9). We do not find any good matches to TEs in any of the large contigs. Thus, transposons in general appear to be filtered out very efficiently by the IES excision process. The tandem and dispersed repeats in the MAC appear to correspond to noninvasive DNA (e.g., the 5S rRNA genes). Taken together, the fact that mobile (and likely invasive) DNA elements are kept out of the MAC, combined with the fact that both tandem and dispersed noninvasive repeats avoid the excision process, indicates strong support for the foreign DNA hypothesis.

In organisms with RIP, since all duplicated DNA is targeted [94], gene diversification by duplication is suppressed. For example, the fraction of all *Neurospora crassa* genes found in paralogous families is only 19%, a value that falls below the overall correlation line between this fraction and total gene number [84]. In addition, very few gene pairs share greater than 80% amino acid sequence identity [84]. Consistent with the foreign DNA hypothesis, we do not see such signs of suppression of gene family diversification in *T. thermophila*. Large numbers of paralogous genes are found in the genome (1,970 gene families including 10,851 predicted proteins) (Table 3). The fraction of genes in such families in *T. thermophila* (39%) is much higher than that seen in *N. crassa*. Although this fraction is not as high as would be predicted from the observed correlation between total number of genes and the fraction found in paralogous families [84], the fraction of gene pairs sharing greater than 80% amino acid identity is much higher than in *N. crassa* and similar to that found in other sequenced eukaryotes.

Since it is possible some of the 1,970 gene families could have originated by duplications that occurred prior to the origin of the IES excision process, it is more useful to examine recent duplications. We searched for such duplications in multiple ways, including the identification of genes duplicated in the *T. thermophila* lineage relative to other lineages for which genomes are available (Table S10) and by searching for pairs of paralogs with very similar sequences. Both of these classes are abundant in *T. thermophila*, further indicating that the IES excision does not significantly affect expansion of gene families of "native" genes. Thus the ciliate

Table 3. Gene Families

Family Size Range	Number of Families	Total Number of Genes	Examples of Families
201 to 500	5	1,525	K^+ channel protein
101 to 200	5	691	Protein kinase; cysteine proteinase; surface antigen
51 to 100	8	522	ABC transporter ABCB/ABCC; cation-transporting ATPase; serine/threonine kinase
21 to 50	37	1,177	Kinesin II; calcium/calmodulin-dependent protein kinase; GTP-binding protein; glutathione *S*-transferases; surface antigen; cytochrome P450; histidine kinase; ABC transporter ABCG; ABC transporter ABCA; dynein heavy chain; carboxypeptidase-like protein; triacylglycerol lipase; oxalate:formate antiporter; metalloproteinase/leishmanolysin-like peptidase; AAA family ATPase; Kazal-type proteinase inhibitor 1; K^+ channel protein; Tlr 5Rp protein; sugar transport protein; protein phosphatase
11 to 20	91	1,292	
6 to 10	195	1,423	
2 to 5	1,629	4,221	

DOI: 10.1371/journal.pbio.0040286.t003

system of targeting invading DNA has significantly different consequences than RIP.

High gene count in *T. thermophila*. The expansion of gene families helps explain the high gene count in *T. thermophila*, which is higher than that of other protists and even surpasses that of some metazoans (Table 4). The duplication events appear to be spread out over evolutionary time with some being ancient and some quite recent. We searched for but did not find evidence for either whole genome or segmental duplications. We do find extensive numbers of tandemly duplicated genes. In total, 1,603 tandem clusters of between two and 15 genes were found, comprising 4,276 total genes; 67% of these clusters are simple gene pairs and 96% contain five or fewer genes. Thus it appears many of the paralogous genes in *T. thermophila* are the results of separate small duplication events.

The high gene count in *T. thermophila* relative to some other single-celled eukaryotes is not simply a reflection of gene family expansions. For example, when recent gene expansions are collapsed into ortholog sets, we find that humans and *T. thermophila* share more orthologs with each other (2,280) than are shared between humans and the yeast *S. cerevisiae* (2,097) or *T. thermophila* and *P. falciparum* (1,325) (Figure 6), despite the sister phyla relationships of animals and fungi on the one hand and ciliates and apicomplexans on the other. We note that this does not mean that humans and *T. thermophila* are overall more similar to each other than either is to species in sister phyla. For example, humans and *S. cerevisiae* do share some processes that evolved in the common ancestor of fungi and animals. In addition, for orthologs found in all eukaryotes, the human and *S. cerevisiae* genes are more similar in sequence to each other than either is to genes from *T. thermophila*. The higher number of orthologs shared between humans and *T. thermophila* is a reflection of both the loss of genes in other eukaryotic lineages and the retention of a variety of ancestral eukaryotic functions by *T. thermophila*. Consistent with this conclusion, there are 874 human genes with orthologs in *T. thermophila* but not *S. cerevisiae*, 58 of which correspond to loci associated with human diseases (Table S12). Thus genome analysis reveals many cases where *T. thermophila* can continue to complement experimental studies of yeast as a model system for eukaryotic (and human) cell biology [13].

Gene Duplication as an Indicator of Important Biological Processes

One motivation for obtaining the genome sequence of an organism is to advance the study of processes already under investigation. Many researchers, including those who have never worked on this species before, have taken advantage of the publicly available data in an effort to achieve this goal (e.g. [24,95–103]). Rather than focus our bioinformatic analysis on these well-studied processes, we decided to search for evidence in the predicted proteome of processes of particular importance to the organism. Our approach was relatively straightforward—we looked for overrepresentations (compared to other eukaryotes) in the lists of paralogous gene families or lineage-specific gene family expansions associated with a variety of processes. This approach was taken for several reasons. First, searches for differences in large gene families are not as biased by annotation errors as searches focused on individual genes. In addition, large gene families clearly contribute to the large number of genes present in *T. thermophila* compared to other single-celled eukaryotes. We note that many of the available genomes of single-celled eukaryotes are of parasites that were

Table 4. Numbers of Protein-Coding Genes in Various Eukaryotes

Species	Predicted Gene Number	Genome Size (Mb)	Genes/Mb
T. thermophila	27,424	104	264
S. cerevisiae	6,561	13	505
S. pombe	4,824	14	345
P. falciparum	5,279	23	230
T. pseudonana	11,242	34	331
D. discoideum	12,500	34	368
D. melanogaster	13,679	180	76
C. elegans	19,971	103	194
A. thaliana	26,207	125	210
Oryza sativa	46,976	466	101
Fugu rupripes	34,312	365	94
Mus musculus	37,854	Approximately 2,500	15
H. sapiens	35,845	Approximately 2,900	12

DOI: 10.1371/journal.pbio.0040286.t004

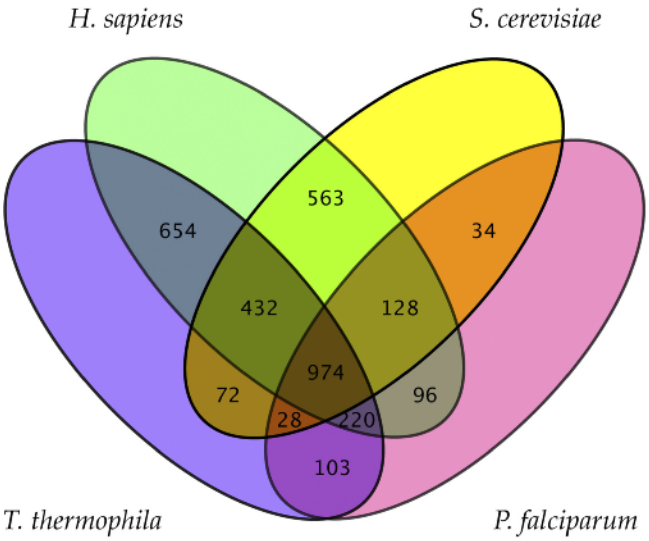

Figure 6. Orthologs Shared among *T. thermophila* and Selected Eukaryotic Genomes

Venn diagram showing orthologs shared among human, the yeast *S. cerevisiae,* the apicomplexan *P. falciparum,* and *T. thermophila*. Lineage-specific gene duplications in each of the organisms were identified and treated as one single gene (or super-ortholog). Pairwise mutual best-hits by BLASTP were then identified as putative orthologs.

DOI: 10.1371/journal.pbio.0040286.g006

selected for sequencing mostly due to their medical relevance and that these are not representative (e.g., many have quite small genomes). Most important, the presence of large gene families and recent gene duplications are likely indications of functional diversity, recent evolutionary innovations, and selective pressures placed on this organism.

Our analysis of paralogous gene families and in particular the recently duplicated members of such families reveals the importance of processes associated with the sensing of and responding to environmental changes. We highlight five such processes here: signal transduction, membrane transport, proteolytic digestion, construction and manipulation of cell shape and movement, and membrane trafficking. These processes are all critical to the free-living heterotrophic lifestyle of this organism. In the following sections, we discuss what the analysis of the genome reveals about these processes in *T. thermophila* with a particular focus on expansions of genes associated with these functions relative to other species.

Signal transduction and the expansions of kinase families. A variety of genes with putative roles in signal transduction were identified in our screens of paralogous genes. Of these, we chose to perform an in depth analysis of the kinases because they are such a diverse family of proteins and because they have been found to have critical roles in sensory and regulatory processes across the tree of life. In total, 1,069 predicted protein kinases (Tables 5 and S11A) were identified in the genome. This corresponds to approximately 3.8% of the predicted proteome, a fraction significantly larger than the approximately 2.3% in fungi, *Drosophila,* and vertebrates [104]. Among these, representatives were found of 54 of the known kinase families and subfamilies [105]. Some families found in a wide diversity of eukaryotes [106] were not detected. This includes the checkpoint kinase CHK1/RAD53, the PI3 kinase–related kinase TRRAP, two cyclin-dependent kinases (CDK7 and CDK8, which may be functionally replaced by the related expanded CDC2 family), and two poorly conserved classes (Bub1 and Haspin) that may have been missed by sequence homology searches. Despite the reported presence of phosphotyrosine in *T. thermophila* [107], no clear members of the tyrosine kinase group could be identified. However, the genome encodes some proteins that might be alternative tyrosine kinases including multiple dual-specificity kinases (e.g., Wee1, Ste7, TTK, and Dyrk) as well as five members of the related TKL group, which may mediate tyrosine phosphorylation in the slime mold *Dictyostelium discoideum* [106]. Twelve kinase classes are found in *T. thermophila* and humans but not yeast, and thus are apparent examples of the retention of ancestral eukaryotic functions discussed above. Several of the genes in these classes have been implicated in the etiology of human disease (Dyrk1A, DNAPK, SGK1, RSK2, Wnk1, and Wnk4) [108].

A key feature of the *T. thermophila* kinome is the expansion of several kinase classes relative to other sequenced organisms (Table 5). The implications of some of these expansions can be predicted based on the known functions of family members. For example, the mitotic kinase families Aurora, CDC2, and PLK are all substantially expanded, perhaps reflecting the additional signaling complexity required by two nuclei that simultaneously engage in very different processes within the same cell cytoplasm. Also expanded are multiple kinases that interact with the microtubule network [109,110] [e.g., Nima-related kinases (NRKs) and the ULK family], possibly reflecting diversification of cytoskeletal systems (discussed more below). Of the kinase families with known functions, the most striking expansion is the presence of 83 histidine protein kinases (HPKs), which are generally involved in transducing signals from the external environment [111]. HPKs are found predominantly in two-component regulatory systems of bacteria, archaea, protists, and plants and are absent from metazoans. Most of the *T. thermophila* HPKs have substrate receiver domains, and many are predicted to be transmembrane receptors.

The full meaning of the kinome diversity in *T. thermophila* is hard to predict as a great deal of the diversification has occurred in classes for which the functions are poorly understood. For example, in many of the known kinase families, the *T. thermophila* proteins are highly diverse in sequence, both relative to those in other species as well as to each other (e.g., see Figure S6). The scope of the diversification in *T. thermophila* is perhaps best seen in the fact that 630 (approximately 60%) of the kinases could not be assigned to any known family or subfamily [105]. Overall, 37 novel classes of kinases and hundreds of unique proteins were identified in this genome. The presence of so many novel kinases and expansions in many known classes of kinases is both an indication of the versatility of the eukaryotic protein kinase domain seen in other lineages [112] and suggestive of a great elaboration of ciliate-specific functions.

Diversification of membrane transport systems. Many of the most greatly expanded *T. thermophila* gene families encode proteins predicted to be involved in membrane transport. Membrane transporters play critical roles in responding to variations in the environment and making use of available resources. We therefore conducted a more thorough analysis of the predicted transporters in this species. Overall, *T. thermophila* possesses a robust and diverse collection of

Table 5. Distribution of Selected Protein Kinase Classes in *T. thermophila* and Other Classified Kinomes

Group	Family	Subfamily	*T. thermophila*	*D. discoideum*	Yeast	Worm	Fly	Human
Human kinases with *T. thermophila* but not yeast homologs								
AGC	MAST		3	5	0	1	2	5
AGC	RSK	RSK	2	0	0	1	1	4
Atypical	PIKK	DNAPK	1	1	0	0	0	1
CMGC	CDK	PITSLRE	1	2	0	2	1	1
CMGC	CDKL		4	0	0	1	1	5
CMGC	Dyrk	PRP4	1	1	0	1	1	1
CMGC	Dyrk	Dyrk1	1	1	0	1	1	2
CMGC	Dyrk	Dyrk2	5	1	0	3	2	3
CMGC	MAPK	p38	2	0	0	3	3	4
CMGC	MAPK	Erk7	3	1	0	1	1	1
Other	TLK		2	0	0	1	1	2
Other	Wnk		2	0	0	1	1	4
Expanded in *T. thermophila*								
Atypical	HistK		83	14	1	0	0	0
Other	ULK		52	2	1	2	3	5
Other	Nek/NRK		39	4	1	4	2	11
Other	Aur		15	1	1	2	2	3
CMGC	CDK	CDC2	11	1	1	2	2	3
CMGC	RCK		8	1	1	1	1	3
CAMK	CAMKL	AMPK	7	1	1	2	1	2
CMGC	MAPK	Erk7	3	1	0	1	1	1
Other	PLK		8	1	1	3	2	4
CAMK	CAMKL	MARK	9	3	1	2	3	4
CMGC	CDKL		4	0	0	1	1	5
STE	Ste20	MST	4	2	1	1	1	2
CMGC	Dyrk	Dyrk2	5	1	0	3	2	3
CMGC	MAPK	Erk	7	1	6	1	1	5
Other	TLK		2	0	0	1	1	2
Eukaryotic "core" kinases not found in *T. thermophila*								
Atypical	PIKK	TRRAP	0	1	1	1	1	1
CAMK	RAD53		0	5	1	2	1	1
CK1	CK1	CK1-D	0	1	1	1	1	2
CMGC	CDK	CDK7	0	1	1	1	1	1
CMGC	CDK	CDK8	0	1	1	1	1	2
Other	Bub		0	1	1	1	2	2
Other	Haspin		0	1	2	13	1	1

Counts are numbers of kinase domains.
Yeast, *S. cerevisiae;* worm, *C. elegans;* fly, *D. melanogaster.*
DOI: 10.1371/journal.pbio.0040286.t005

predicted membrane transport systems (Tables 6 and S11B). Comparison to other eukaryotes [113] reveals some interesting differences in terms of both classes of transporters and predicted substrates being moved. For example, *T. thermophila* has more representatives in each of the four major families than do humans. In addition, it encodes a much higher number of transporters in the ABC superfamily, voltage-gated ion channels (VICs), and P-type ATPases than any other sequenced eukaryotic species (Table 6) including the other free-living protists, the diatom *Thalassiosira pseudonana,* and the slime mold *D. discoideum*. Regarding substrates, an extremely extensive set of transporters likely specific for inorganic cations has been identified (Table 6). Most of these are channel-type transporters and cation-transporting P-type ATPases. Interestingly, despite the apparent massive amplification of cation transporters, *T. thermophila* has a very limited repertoire of transporters for inorganic anions: only one member each for sulfate, phosphate, arsenite, and chromate ion were identified, and there are no predicted anion channels. The reason for the difference in the amplification of cation versus anion transporters is unclear.

As with kinases, some of the most interesting properties are revealed by examination of the lineage-specific duplications of transporters. The recent clusters include K^+ channel proteins (285 members), ABC transporters (152 members), cation-transporting ATPases (59 members), K^+ channel beta subunit proteins (22 members), oxalate:formate antiporters (24 members), sugar transporters (22 members), and phospholipid-transporting ATPases (20 members). The expansion of the K^+ channel proteins, which are VIC-type transporters, was particularly large and was pursued further.

In total, 308 VIC-type K^+-selective channels have been predicted, many more than in any other sequenced species and over three times as many as identified in humans (89). A multigene family of potassium ion channels has also been identified in *P. tetraurelia* [114] and thus may be a general characteristic of some ciliates. Some lines of evidence suggest that this expansion in ciliates could be adaptive. First, K^+ channels control the passive permeation of K^+ across the membrane, which is essential for ciliary motility [115]. Second, a novel adenylyl cyclase with a putative N-terminal K^+ ion channel regulates the formation of the universal second messenger cAMP in ciliates and apicomplexans

Table 6. Comparison of the Numbers of Membrane Transporters in *T. thermophila* and Other Eukaryotes by Family and Predicted Substrate

Species	Family					Predicted Substrate							Total	Percent of ORFs
	ABC	MFS	VIC	P-ATPase	Other	Inorganic Cations	Inorganic Anions	Carbon Compounds	Amino Acids and Derivatives	Bases and Derivatives	Vitamins and Cofactors	Drugs, Toxins, and Macromolecules	Unknown	
T. thermophila	161	125	332	91	231	485 (51.6%)	15 (1.6%)	77 (8.2%)	49 (5.2%)	26 (2.8%)	23 (2.4%)	155 (16.5%)	110 (11.7%)	940	3.4%
E. histolytica	18	4	1	19	57	27 (27.3%)	11 (11.1%)	6 (6.1%)	10 (10.1%)	2 (2%)	3 (3%)	31 (31.3%)	9 (9.1%)	99	1.0%
D. discoideum	61	27	3	24	135	54 (21.6%)	23 (9.2%)	22 (8.8%)	27 (10.8%)	7 (2.8%)	9 (3.6%)	61 (24.4%)	50 (20%)	250	1.8%
T. pseudonana	55	42	22	22	271	103 (25%)	53 (12.9%)	42 (10.2%)	56 (13.6%)	11 (2.7%)	27 (6.6%)	83 (20.1%)	43 (10.4%)	412	3.6%
C. parvum	13	8	2	9	43	17 (22.7%)	4 (5.3%)	7 (9.3%)	11 (14.7%)	2 (2.7%)	11 (14.7%)	11 (14.7%)	12 (16%)	75	2.2%
P. falciparum	14	15	1	11	47	25 (28.4%)	6 (6.8%)	9 (10.2%)	3 (3.4%)	4 (4.5%)	6 (6.8%)	14 (15.9%)	21 (23.9%)	88	1.7%
Encephalitozoon cuniculi	11	2	0	4	26	11 (25.6%)	2 (4.7%)	2 (4.7%)	7 (16.3%)	4 (9.3%)	13 (30.2%)	4 (9.3%)	0 (0%)	43	2.2%
N. crassa	31	141	2	19	153	63 (18.2%)	18 (5.2%)	83 (24%)	28 (8.1%)	7 (2%)	3 (0.9%)	85 (24.6%)	44 (12.7%)	346	3.4%
S. cerevisiae	24	85	2	16	176	59 (19.5%)	21 (6.9%)	63 (20.8%)	38 (12.5%)	11 (3.6%)	8 (2.6%)	59 (19.5%)	39 (12.9%)	303	4.8%
S. pombe	9	58	1	13	107	45 (23.9%)	13 (6.9%)	22 (11.7%)	26 (13.8%)	5 (2.7%)	3 (1.6%)	35 (18.6%)	36 (19.1%)	188	3.8%
A. thaliana	108	90	35	46	643	245 (26.6%)	95 (10.3%)	101 (11%)	119 (12.9%)	38 (4.1%)	40 (4.3%)	151 (16.4%)	149 (16.2%)	922	3.5%
C. elegans	48	134	63	22	389	181 (27.6%)	108 (16.5%)	51 (7.8%)	122 (18.6%)	23 (3.5%)	28 (4.3%)	37 (5.6%)	106 (16.2%)	656	4.0%
D. melanogaster	51	136	31	19	361	142 (23.7%)	77 (12.9%)	84 (14%)	105 (17.6%)	9 (1.5%)	14 (2.3%)	69 (11.5%)	99 (16.6%)	598	3.2%
H. sapiens	47	81	89	32	521	261 (33.9%)	82 (10.6%)	86 (11.2%)	94 (12.2%)	13 (1.7%)	19 (2.5%)	75 (9.7%)	142 (18.4%)	770	2.8%

[a]Percent of total transporters are indicated in parentheses.

ABC, ATP-binding cassette; MFS, major facilitator superfamily; VIC, voltage-gated ion channels; P-ATPase, P-type ATPase.

DOI: 10.1371/journal.pbio.0040286.t006

[116,117], which could assist in responding to sudden changes of the ionic environment. *T. thermophila* encodes six homologs of this adenylate cyclase/K^+ transporter, whereas the parasitic apicomplexans *P. falciparum* and *Cryptosporidium parvum* encode only one each.

The robust transporter systems present are likely a reflection of *T. thermophila's* behavioral and physiological versatility as a free-living single-celled organism and its exposure to a wide range of different substrates in its natural environment. Examination of the specific types of expansions suggests that functions associated with transport of K^+ and other cations have been greatly diversified. Thus such functions may play a role in many of the unique aspects of the biology of this species and ciliates in general.

Proteolytic processing. *T. thermophila* is a voracious predator and thus might be expected to have a wide diversity of proteolytic enzymes. Analysis of the predicted proteins in *T. thermophila* reveals some conflicting results relating to this idea. On the one hand, many of the largest clusters of lineage-specific duplications are of proteases (e.g., papain, leishmanolysin). On the other hand, the total number of proteases identified (480) is relatively low in terms of the fraction of the proteome (1.7%) compared to other model organisms that have been sequenced and annotated [118–120]. The conflict is most likely a reflection of the diversity of physiological processes in which proteases function [121]. Thus we examined the subclassification of types of proteases present in more detail.

Using the Merops protease nomenclature, which is based on intrinsic evolutionary and structural relationships [119] the *T. thermophila* proteases were divided into five catalytic classes and 40 families. These are: 43 aspartic proteases belonging to two families, 211 cysteine proteases belonging to 11 families, 139 metalloproteases belonging to 14 families, 73 serine proteases belonging to 12 families, and 14 threonine proteases belonging to the T1 family (Tables 7 and S11C). Some unique features of *T. thermophila* can be seen by comparison to *P. falciparum* which is the most closely related sequenced species to have a detailed analysis of its proteases published [122]. Twenty-one protease families are present in both genomes. For example, the highly conserved threonine proteases and the ubiquitin carboxyl-terminal hydrolase families (C12 and C19) reflect the crucial role of the ATP-dependent ubiquitin-proteasome system, which has been implicated in cell-cycle control and stress response [123]. Nineteen protease families are present in *T. thermophila* but not *P. falciparum.* One of these includes leishmanolysin (M8), originally identified in the kinetoplastid parasite *Leishmania major* and thought to be involved in processing surface proteins [124–126]. This family is greatly expanded (to 48 members, including 15 in a tandem array) in *T. thermophila* and suggests that surface protein processes may be important here, although the functions of leishmanolysin-related proteases in nonkinetoplastid eukaryotes remain unclear. The carboxypeptidase A (M14) and carboxypeptidase Y (S10) families are expanded to 28 and 25 members, respectively, in *T. thermophila,* which may reflect numerous and diverse functions. Only four protease families present in *P. falciparum* are not found in *T. thermophila.* Among these are metacaspase (C14), an ancestral type of caspase that is characteristic of apoptosis or apoptosis-like signal transduction pathways [127].

The largest clusters of expanded proteases in *T. thermophila*

Table 7. Protease Complements in *T. thermophila* and Other Model Organisms

Organism	Catalytic Class					Total	Percentage of the Genome[a]
	Aspartic	Cysteine	Metallo	Serine	Threonine		
T. thermophila	43 (9.0%)[b]	211 (44.0%)	139 (28.9%)	73 (15.2%)	14 (2.9%)	480	1.7
P. falciparum[c]	10 (10.5%)	33 (34.7%)	21 (22.1%)	16 (16.9%)	15 (15.8%)	95	1.8
S. cerevisiae	14 (9.5%)	43 (29.0%)	49 (33.1%)	26 (17.6%)	16 (10.8%)	148	2.4
A. thaliana	203 (24.5%)	154 (18.6%)	110 (13.2%)	326 (39.3%)	37 (4.4%)	830	2.7
C. elegans	27 (6.0%)	114 (25.3%)	180 (40.0%)	105 (23.3%)	24 (5.3%)	450	2.2
D. melanogaster	46 (6.6%)	80 (11.4%)	191 (27.2%)	351 (50.1%)	33 (4.7%)	701	5.1
M. musculus	91 (11.7%)	162 (20.9%)	205 (26.4%)	285 (36.7%)	33 (4.3%)	776	2.8
H. sapiens	312 (31.6%)	167 (16.9%)	223 (22.6%)	247 (25.1%)	37 (3.8%)	986	4.1
E. coli	12 (6.2%)	30 (15.5%)	60 (31.1%)	87 (45.1%)	4 (2.1%)	193	3.9
Methanococcus jannaschii	2 (5.3%)	11 (29.0%)	17 (44.7%)	5 (13.1%)	3 (7.9%)	38	2.6

[a]The percentage of the whole genome that encodes putative proteases.
[b]Percentage of individual catalytic class in the protease complement is included in parentheses.
[c]The distribution of proteases in *P. falciparum* is based on Wu et al. [122], and the distributions in the other model organisms are based on the results published in the Merops database Release 7.00.
DOI: 10.1371/journal.pbio.0040286.t007

are all cysteine proteases, which comprise 44% of the total protease complement. The two most prominent families from this class are the papain family (C1), which is the most abundant and complex family, with 114 members, and the ubiquitin carboxyl-terminal hydrolase 2 family (UCH2, C19) with 47 members. It is possible that the biochemical activity among the paralogs within these families is conserved but that they are used in different parts of the cell (or outside the cell) or in different developmental stages in *T. thermophila*.

Cytoskeletal components and regulators. Ciliates have highly complex cytoskeletal architecture [128] with highly polarized cell types which assemble 18 types of microtubular organelles in specific locations along the anteroposterior and dorsoventral axis. We therefore sought to determine whether this diversity was reflected in the genome. As with the protease analysis described above, initial comparisons of the number of particular types of cytoskeletal and microtubule-associated proteins was somewhat ambiguous (the numbers for humans and *T. thermophila* are shown in Tables 8 and S11D). For example, although kinesin and dynein motors as well as kinases associated with microtubules appear to be expanded, structural components of the cilia and participants in the intraflagellar transport pathway are not. In addition, some cytoskeletal protein types are apparently absent from *T. thermophila;* these include intermediate filament proteins (including nuclear lamins) as already suggested by biochemical studies [129], some microtubule-associated proteins (MAP2, MAP4, and Tau, for which no nonanimal eukaryotic homologs have been found) and some actin-binding proteins (e.g., α-actinin). To better understand what role genes involved in microtubule and cytoskeletal functions might have played in the diversification of this species, we focused analysis on some of the genes with apparent expansions: tubulins, dyneins, and regulatory proteins.

Tubulins. Tubulins are the key structural components of microtubules and they come in many forms in eukaryotes [130]. In the *T. thermophila* genome, phylogenetic analysis of tubulin homologs (Figure 7) reveals the presence of one or two genes, each within the essential alpha (α), beta (β), and gamma (γ) subfamilies (as reported previously [131–133]) and one in each of the delta (δ), epsilon (ε), and eta (η), which are found in organisms that possess centrioles/basal bodies [134–136]. In addition, *T. thermophila* encodes noncanonical tubulin homologs that can be divided into two categories. In the first category are genes that are most similar to the canonical α- or β-tubulins. These nine genes (three α-like and six β-like) lack characteristic motifs for the tail domain post-translational modifications (polyglutamylation and polyglycylation) that are essential to the function of their canonical counterparts [137–139]. Three of the β-like genes (*BLT1*/TTHERM__01104960, TTHERM__01104970, and TTHERM__01104980) form a tandem cluster with intergenic intervals of less than 2 kb. We hypothesize that these genes function, perhaps redundantly, in formation or function of some of the many highly specialized microtubule systems of *T. thermophila* cells. Experimental analysis of *BLT1*, a β-like tubulin, indicated that its product localizes to a small subset of microtubules and is not incorporated into growing ciliary axonemes (K. Clark and M. Gorovsky, unpublished data). Genetic deletion of this gene or of the α-like gene TTHERM__00647130 did not yield an obvious phenotype (R. Xie and M. A. Gorovsky, unpublished data).

The second category of noncanonical tubulin homologs consists of three novel proteins (TTHERM__00550910, TTHERM__01001250, and TTHERM__01001260) that fall into a clade with *P. tetraurelia* iota tubulin. Two of these (TTHERM__01001250 and TTHERM__01001260) are closely related to each other (Figure 7) and closely linked in the genome and thus likely arose by a recent tandem duplication. The functions of these genes are unknown, but because they are, so far, unique to ciliates, they might be responsible for microtubule functions specific to this phylum.

Dyneins. Dyneins, which were first discovered in *Tetrahymena* [140], are molecular motors that translocate along microtubule tracks, a process critical to many activities in *T. thermophila* including ciliary beating, karyokinesis, MAC division, cortical organization, and phagocytosis. Many of these activities are critical for sensing and responding to changes in the environment. Each dynein complex consists of one, two, or three heavy chains (containing the motor

Table 8. Numbers of Loci Encoding Selected Types of Cytoskeletal Genes in *T. thermophila* and *H. sapiens*

Protein Type	*T. thermophila*	*H. sapiens*
Actin-related	14	19
Actin-binding proteins		
Profilin	1	2
α-Actinin	0	4
Fascin	0	3
Cofilin	1	3
Gelsolin	0	2
CapZ	1	3
Tropomodulin	0	4
Paxillin	1	4
Fimbrin	1	2
Intermediate filaments		
Desmin	0	1
Vimentin	0	1
Keratin	0	8
Lamin (A/C, B)	0	3
Tubulins		
α–tubulin	1	9
α–tubulin-like	3	0
β–tubulin	2	9
β-tubulin-like	6	0
γ–tubulin	1	2
ε–tubulin	1	1
δ-tubulin	1	1
η-tubulin	1	0
κ-tubulin	3	0
Microtubule-associated proteins		
MAP1A	0	1
MAP1B	0	1
MAP2	0	1
MAP4	0	1
Tau	0	1
TPX2	1	1
XMAP215	2	1
EB1	7	3
Centrin	6	3
Pericentrin	0	2
Katanin (p60)	2	2
Motor proteins		
Kinesin motor chain	78	48
Dynein motor chain	25	46
Myosin motor chain	13	22
Tubulin-modifying enzymes		
Tubulin deacetylase HDAC6	2	1
Tubulin tyrosine ligase-like	50	14
Intraflagellar transport (IFT) components		
IFT20	1	1
IFT52	1	1
IFT57	1	1
IFT71	1	1
IFT81	1	1
IFT88	2	1
IFT140	1	1
IFT172	1	1
Structural components of cilia and flagella		
Radial spoke protein 4/6	3	2
Radial spoke protein 2	3	1
PF16	1	1
PF20	1	1
Cytoskeleton-associated serine-threonine kinases		
NIMA-related kinase (NRK)	39	11
Aurora kinase	16	3
Polo kinase	8	4

DOI: 10.1371/journal.pbio.0040286.t008

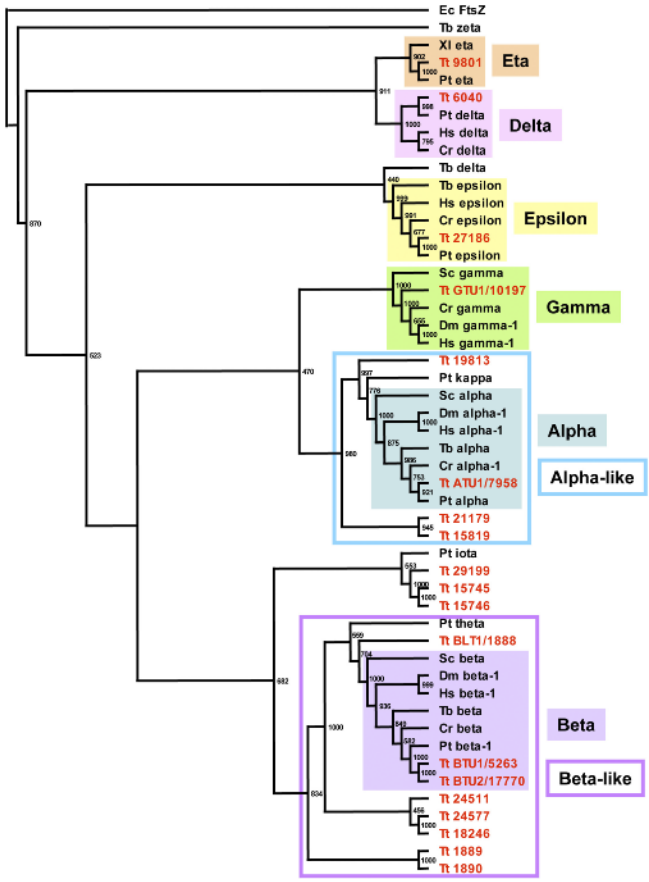

Figure 7. Tubulin Gene Diversity in *T. thermophila*

The figure shows a neighbor-joining tree built from a clustalX alignment. Species abbreviations: Hs, *H. sapiens;* Dm, *D. melaogaster;* Sc, *S. cerevisiae;* Tt, *T. thermophila;* Pt, *P. tetraurelia;* Cr, *C. reinhardtii;* Tb, *T. brucei;* Ec, *E. coli;* Xl, *Xenopus laevis.* A prokaryotic tubulin ortholog, *Escherichia coli* FtsZ, was used as the outgroup.

DOI: 10.1371/journal.pbio.0040286.g007

activity) and specific combinations of smaller subunits, including intermediate, light-intermediate, and light chains, which regulate motor activity and the tethering of dynein to its molecular cargo [141–143]. In organisms with cilia or flagella, there are multiple isoforms of dyneins, including the axonemal outer arm dyneins, the axonemal inner arm dyneins, and nonaxonemal or "cytoplasmic" dyneins. Each is specialized in its intracellular location and the cellular task it performs [144].

In total we identified 21 light chains, five intermediate chains, two light-intermediate chains, and 25 heavy chains (Table S13). The expression of each gene, as well as the exon/intron structures of most, was confirmed by RT-PCR and, if necessary, sequencing of the RT-PCR product. For the most part, the families of *T. thermophila* dynein subunits appear to be similar to those of other model organisms; however, there are some interesting differences. *T. thermophila* light chains LC3A and 3B are most similar to the green alga *Chlamydomonas reinhardtii*'s LC3 and LC5 [145]. These proteins belong to the larger family of thioredoxin-related proteins, and, without biochemical evidence identifying one or both of the proteins as part of a dynein complex, it may be premature to label these as dynein components. Light chain LC4 belongs to the calmodulin-related family of proteins and may regulate

calcium-dependent ciliary reversal. *T. thermophila* expresses two LC4 genes, perhaps providing alternative or additional ways to control ciliary motility compared to species that express only one. In other systems, LC8 is associated with several different dynein and nondynein complexes, and *T. thermophila* expresses one canonical LC8 as well as five divergent LC8-like genes, with unknown functions.

Perhaps the most interesting revelation is that *T. thermophila* expresses 25 dynein heavy chains. These include the 14 DYH genes previously described [146,147] and 11 new ones, all of which appear to be axonemal. The complexity of the DYH family may represent a mechanism by which the organism can fine-tune ciliary activity, produce specialized cilia (e.g., oral and posterior cilia), and/or generate large numbers of new cilia quickly. Along these lines, there has also been an expansion in other motor proteins. For example, there are 78 kinesins, more than in any other sequenced organism ([101] and Table 8). In addition, although there are fewer myosins than in humans (13 versus 22), 12 of 13 of the *T. thermophila* genes comprise a single novel myosin class not found in other organisms [102,148].

Regulation of microtubules and microtubule-associated processes. Among the expanded genes in *T. thermophila* are a variety implicated in the regulation of microtubules or microtubule-associated processes. One example is the tubulin tyrosine ligase-like domain proteins of which multiple members have been identified as enzymes responsible for polyglutamylation of either α- or β-tubulin [149]. *T. thermophila* encodes 50 tubulin tyrosine ligase-like proteins compared with 14 in human. Another example is the NRK family of protein kinases which, as mentioned above, has undergone a large expansion in *T. thermophila*. NRKs are often found associated with microtubular organelles [150] such as centrioles, basal bodies, and flagella and play multiple roles, including the regulation of centrosome maturation [151] and flagellar excision [152]. We identified 39 NRKs in *T. thermophila,* roughly three times the number of such loci in humans. Phylogenetic and functional analyses have suggested that this diversification has adapted the members of this family for distinct subcellular localizations and cytoskeletal roles [103]. Thus, such gene expansions could allow differentially targeted protein isoforms to regulate the function of the same organelle type in different locations or generate different properties of the same structural building materials (e.g., microtubules), which are used as frameworks to build different types of organelles.

Secretory pathways and membrane trafficking. Besides the conventional organelles, *T. thermophila* maintains several more specialized membrane-bound compartments, including alveoli (shared with other alveolates), a contractile vacuole (found in many protists), and separate, functionally distinct macronuclei and micronuclei [128]. It also has multiple pathways for plasma membrane internalization, as well as both constitutive and regulated exocytosis [128,153]. The sorting and trafficking of membrane components are critical functions for all these activities. Analysis of the genome reveals homologs of many of the key proteins known from other eukaryotes to be involved in vesicle formation and fusion, including all major classes of coat proteins (Table S14). One interesting finding that came from genome analysis is that *T. thermophila* encodes eight dynamin-related proteins, more than most other sequenced unicellular eukaryotes, and two of them, Drp1p and Drp2p, have evolved a new function in endocytosis [96] (A. Rahaman and A. P. Turkewitz, unpublished data). Furthermore, phylogenetic analysis indicated that the recruitment of dynamin to a role in endocytosis occurred independently by convergent evolution in the animal and ciliate lineages [96].

The diversification of membrane trafficking is more apparent in regard to Rab proteins, which are small monomeric GTPases that regulate membrane fusion and fission events. *T. thermophila,* with 69 Rabs (Table S15), has a number more along the lines of humans (which have 60) than many single-celled species, such as *Saccharomyces cerevisiae*, which has 11 [154] and *Trypanosoma brucei,* which has 16 [155]. Based on localization and functional studies, including comparisons between yeast and humans [156], Rabs have been divided into eight groups [157]. Phylogenetic analysis (Figure S7) indicates that *T. thermophila* encodes representatives of all but groups IV and VII, which are involved in late endocytosis and Golgi transport, respectively. For group VII this appears to reflect a lineage-specific loss, since the genomes of both *T. brucei* and *Entamoeba histolytica* have several homologs in this group. Two *T. thermophila* Rabs appear homologous to Rab28 and Rab32, which have not been assigned to any of these groups; Rab32 was previously thought to be restricted to mammalian lineages. Rab groups II and V, involved in endocytosis, are especially large in *T. thermophila* and include several Rab2, Rab4, and Rab11 homologs in group II. This may reflect the intricacy of maintaining at least two major pathways of membrane internalization. Additionally, 29 Rabs in *T. thermophila* fail to cluster with any of the Rab groups found more widely among eukaryotes. Within this group, 20 cluster into three clades, designated *Tetrahymena* clades I, II, and III in Figure S7, which may represent ciliate-specific radiations. The remaining nine are very divergent and may represent very ancient duplication events and/or changes related to recruitment for novel function. Because unambiguous alignment among such divergent Rabs is difficult, their relationships will become clearer as additional related genomes are sequenced.

Recently, large numbers of Rabs have been found in a variety of amoeboid protists including *D. discoideum, E. histolytica* [158], and the parabasalid *Trichomonas vaginalis* [159]. The diversification in these species was proposed to relate to their amoeboid lifestyle [159]. However, the presence of significant diversification in *T. thermophila* suggests that different protist lifestyles may be accompanied by their own brand of significant Rab diversification.

Tetrahymena Genome Database

An integral part of the effort to make the genomic resources and analyses described above widely available to researchers working with *T. thermophila* and other organisms has been the creation of the *Tetrahymena* Genome Database (TGD; http://www.ciliate.org), a Web-accessible resource on the genetics and genomics of *T. thermophila*. TGD provides information about the *T. thermophila* MAC genome, its genes and gene products, facts about the ciliate scientific community, and tools for querying the genome and collected scientific literature. TGD was created using the database environment developed for the *Saccharomyces* Genome Database and software tools contributed to the Generic Model Organism Database (GMOD) project.

Information from the published literature on *T. thermophila* is distilled in multiple ways. Results from published studies of *T. thermophila* genes are curated and provided, including community-approved gene names, other nonstandard aliases, nucleotide and amino acid sequences, and literature citations. In addition, free-text descriptions are associated with predicted gene models, and full-text searching is provided using Textpresso [160]. To enable intra- and cross-species comparisons, when information on characterized genes is curated, TGD staff members capture aspects of a gene product's biology (i.e., molecular function, biological role, and cellular localization) using terms from the Gene Ontology (http://www.geneontology.org). This is complemented by automated functional annotation of all predicted genes. Other resources include tools for searching the annotation by keywords, similarity searching using BLAST and BLAT, Gbrowse-based genome visualization [161], information about *Tetrahymena* research laboratories, links to other ciliate-related resources, and various tutorials. The TGD staff is always available to help individual researchers by answering questions, finding information, and generating datasets specific to their needs.

Conclusions and Future Plans

In sequencing and assembling the *T. thermophila* MAC genome, there were many anticipated major challenges not commonly seen in eukaryotic genome projects. Overall, however, the assemblies are remarkably accurate and represent excellent coverage of the genome. This is likely in large part due to low levels of repetitive DNA, one of the features of the MAC genome that initially led us to select it for sequencing. The sequence data in our current assemblies are certainly complete enough for detailed analyses of the predicted biology of this species as we have reported here and others have shown. In addition, the genome sequence is already being used in many functional genomic studies taking advantage of the powerful experimental tools available. Along these lines, it will be of great value to do comparative analyses with the genome sequences of other ciliates such as *P. tetraurelia* and *Oxytricha trifallax,* which are in progress.

One of our main goals is to obtain a complete sequence of the MAC genome, and there are still some challenges left to its achievement. Since we were unable to obtain quality sequence data from large insert clones, any region of the MAC genome containing significant amounts of repetitive DNA would not have assembled well. To overcome this pitfall we are now using HAPPY mapping [162] as an alternative approach to obtaining such linking information. Also, it is known that at least the ends of at least two MAC chromosomes present immediately following conjugation disappear during subsequent vegetative growth, perhaps an indication that these chromosomes are incapable of long-term maintenance [41]. As expected, we do not find sequences corresponding to these ends in our database. Thus alternative methods will be required to obtain the sequences of these regions and any others lost during early vegetative growth. Despite these challenges, all the evidence suggests that it will be possible to close the entire MAC genome.

Of course, the entire MAC genome alone does not provide us with a complete picture of the *T. thermophila* genome. Sequencing the MIC genome will be more challenging due to the greater abundance of repetitive DNA. However, we will be able to use the MAC genome as a scaffold and thus in a way MIC sequencing will be equivalent to genome closure rather than an independent project. We have already begun in this area by determining the sequence adjacent to MIC Cbs junctions and mapping these to MAC assemblies as well as the reverse—using MAC telomere-adjacent sequences to pull out MIC Cbs-flanking regions [24,41].

Having a MIC sequence and mapping the MIC to the MAC will be useful in understanding many aspects of *T. thermophila* biology that we cannot study through the MAC. These include centromere function, MIC telomere features, and the extent to which the MAC and MIC in *T. thermophila* and other ciliates are the equivalent of somatic and germ cells. Perhaps most important, having both genomes will allow detailed analyses of the genome-wide DNA rearrangement process. It is only by having both genome sequences that we can fully understand the biology of this fascinating species.

Materials and Methods

Cell growth, DNA isolation, and library construction. *T. thermophila* cell lines currently in laboratory use were first isolated from the wild in the 1950s [163] and were maintained by serial passage and inbreeding for over 16 y before viable freezing methods were developed. Strain SB210 [164] is the end result of about 25 sexual reorganizations in laboratory culture, including a series of sexual inbreedings by the equivalent of brother-sister matings giving rise to the inbred strain B genetic background [165]. Following the final conjugation, a thoroughly assorted cell line was isolated after at least three serial single-cell isolations (SCIs). The last SCI was approximately 150 fissions after conjugation. These serial SCIs provided abundant opportunity to isolate a cell line that had become pure for most of the MAC developmental diversity but not necessarily all because assortment brings about a stochastic, exponential decay in diversity. The chosen cell line was then subjected to a genomic exclusion cross [166], which generates a whole-genome homozygous MIC but does not generate a new MAC. At least one additional SCI occurred at this step, after which this cell line was frozen. As needed, frozen stocks were replenished following a minimal number of vegetative fissions. The strain has been deposited in the *Tetrahymena* Stock Center at Cornell University as suggested [167].

A culture was started from a fresh thaw of strain SB210. Purified macronuclei were prepared by differential sedimentation, and DNA was extracted from the purified macronuclei as described [168]. The preparation was checked by Southern blot hybridization to verify that the level of contamination with MIC DNA was low. Genomic libraries were prepared as described [169]. DNA was randomly sheared, end-polished with consecutive polynucleotide kinase and T4 DNA polymerase treatments, and size-selected by electrophoresis in 1% low-melting-point agarose. After ligation to BstXI adapters (Invitrogen, Carlsbad, California, United States; catalog No. N408–18), DNA was purified by three rounds of gel electrophoresis to remove excess adapters, and the fragments, now with 3′-CACA overhangs, were inserted into BstXI-linearized plasmid vector (pHOS2, a medium-copy pBR322 derivative) with 3′-TGTG overhangs. Libraries with average sizes of inserts were constructed: 1.8, 2.5, 3.5, 5.0, and 8.5 kb (Table S1). Libraries with larger insert sizes were unstable, presumably due to the high AT content in the genomic DNA.

Sequencing was done from paired-ends primarily at the J. Craig Venter Science Foundation Joint Technology Center. Possible contaminating sequences from other projects have been filtered out using BLASTN searches against all other genome projects conducted at the same time at TIGR and the Joint Technology Center. Whole genome assemblies were performed using the Celera Assembler [37] with modifications implemented by researchers at the J. Craig Venter Science Foundation and TIGR. Sequence reads corresponding to the mitochondrial and rDNA chromosomes were identified using the latest version of the MUMmer program [170] and comparison to the published sequences.

Linking open ends of assembled scaffolds to telomeres. The initial assembly contained 85 telomere-capped scaffold ends. However, these ends correspond to a minority of the total number of non–rDNA telomere–containing sequence reads, which we estimate to be

4,058. Computational and experimental methods were used to identify and confirm scaffold ends that were very close to a telomere, marking the end of a chromosome.

One method matched read-mates of telomere-containing reads (Tel-reads) that the assembly program failed to incorporate into scaffolds. These were identified by searching the sequence read database for exact matches to a 12-mer encompassing two telomeric repeats (GGGGTTGGGGTT). Read-mates were identified for 95% of the Tel-reads. Two internal 40-nt tags were extracted from each Tel-read mate and tested for at least one exact match with the terminal 5 kb of every scaffold (or the entire scaffold if less than 10 kb long). After clustering the matches, a nonredundant list of Tel-linked scaffold ends was generated.

A second method matched previously identified MIC DNA sequences flanking cloned Cbs junctions to scaffold ends (see Figure 2). Telomeres are added within 30 bp of the Cbs element. Thus, if Cbs-adjacent sequence from MIC DNA can be aligned with a MAC scaffold end, the end can be inferred to be telomere-linked. BLASTN searches were carried out with the "no filter" option because very AT-rich sequence was being compared.

A third method involved PCR walking from scaffold open ends to telomeres. Primers designed from scaffold ends were used in combination with the generic 14-nt telomere primer, 5′-CCCCAACCCCAACC-3′. The authenticity of each PCR product was confirmed by sequencing.

Cloning and sequencing RAPDs and sizing their associated MAC chromosomes. Conditions and reagents for RAPD PCR were as in [171]. The 10-mer primers were from Operon Technologies. The polymorphic RAPD PCR products were size-fractionated by electrophoresis in a 1.5% agarose gel. Polymorphic bands were excised and the DNA was extracted with a QIAquick gel extraction kit (Qiagen, Chatsworth, California, United States). The DNA was reamplified using the same PCR conditions and primer combination initially used to detect the polymorphism. Amplified fragments were cloned into the pCR2.1-TOPO vector (Invitrogen) according to the manufacturer's directions. Insert-containing clones, identified as white colonies, were screened for insert size by colony PCR as in [172]. The authenticity of each correctly sized insert was confirmed by hybridization to a Southern blot of RAPD products from a panel of ten *Tetrahymena* strains in which the alleles of the RAPD locus were meiotically segregating [40].

Plasmid DNA was isolated using a QIAprep Miniprep kit (Qiagen, Valencia, California, United States), and inserts were sequenced using the Big Dye Terminator Cycle-Sequencing-Ready Reaction kit (PE Applied Biosystems, Foster City, California, United States). Nucleotide sequences were determined using an ABI 310 Genetic Analyzer. Insert sequences were then searched against the assemblies using BLASTN.

High-molecular-weight DNA was prepared by embedding live cells from strain SB210 in agarose plugs and lysing them using a modification of Birren and Lai [173]. The DNA plugs were inserted into the wells of a 1% Pulsed Field Certified Agarose gel (Bio-Rad, Hercules, California, United States) in 1× TAE buffer. Preliminary sizing of MAC chromosomes was obtained from gels run using the following conditions: 30 h at 6 V/cm with a 60- to 120-s switch time ramp at an included angle of 120°, 1× TAE recirculated at 10 °C. Running conditions were varied when the above conditions did not provide adequate resolution in the size range of a particular MAC chromosome (E. P. Hamilton, unpublished data). The DNA in the gel was acid-depurinated, neutralized, and transferred to a positively charged nylon membrane by downward alkaline transfer (CHEF-DR III Instruction Manual and Applications Guide; Bio-Rad). After blotting, the DNA was crosslinked to the membrane using a Bio-Rad GS Gene Linker. ^{32}P-labeled probes were made from the PCR products obtained from each RAPD clone. Methods for making probes, Southern hybridization, and autoradiography were as in [40].

cDNA library construction and sequencing. cDNA libraries were generated from cells in either the conjugative or vegetative stages of the life cycle. For the conjugative library, cells from a mating between strains CU428 and B2086 were harvested at 3, 6, and 10 h after mixing, and RNA was purified using TRIzol. PolyA+ RNA was isolated and cDNA was generated by Amplicon Express (Pullman, Washington, United States). Inserts were cloned into EcoRI and XhoI sites in pBluescript IISK+ (Stratagene, La Jolla, California, United States) and had an average size of 1.4 kb. Clones were picked at random and sequenced from the 5′ end of the transcript using the T3 primer. For the vegetative library, which was made by DNA Technologies (Gaithersburg, Maryland, United States), CU428 cells were harvested in exponential growth and RNA was purified using TRIzol. PolyA+ mRNA was isolated using oligo(dT) cellulose, cDNA was generated, and inserts were cloned into the EcoRV and NotI sites of the pcDNA3.1(+) vector (Invitrogen). Clones were picked at random and sequenced from the 5′ end using the custom pcDNA(–48) primer. All sequences were submitted to the dbEST division of GenBank, to the Taxonomically Broad EST Database (TBestDB) at http://tbestdb.bcm.umontreal.ca/searches/login.php, and to TIGR's *Tetrahymena* Gene Index at http://www.tigr.org/tigr-scripts/tgi/T__index.cgi?species=t__thermophila. Subsequent analyses used comparisons of the conjugative sequences with all vegetative sequences including those in GenBank not generated at TIGR.

Functional ncRNA analysis. Most ncRNA annotations (Table S6) were generated using covariance model (CM) scans [174]. Transfer RNA annotations are those provided by the CM-based tRNAscanSE program [175] run with default parameters. Most other scans were based on CMs defined by the Rfam database [176,177] (release 7.0, March 2005; 503 families). With a few exceptions, we used rigorous filters [178] built from the Rfam models to identify exactly those sequences that match the Rfam models with scores at or above Rfam's family-specific "gathering" cutoff. One exception was RF00005 (tRNA), as mentioned above. Another exception was RF00012, the U3 small nucleolar RNA, for which the Rfam model found no hits. Instead, we manually added one known *Tetrahymena* U3 sequence [179] to the Rfam seed alignment, built a CM from it, and rescanned the genome, finding the four U3 sequences reported in Table S6C. The third class of exceptions consisted of the 44 Rfam families using the "local alignment" feature of CMs. These families were scanned using ML-heuristic filters [180], with a scan threshold chosen for each such family such that approximately 1% of the genome was scored by the CM. This setting generally shows good sensitivity but is not guaranteed to find all sequences that match the Rfam model, unlike the rigorous scans above. Hits against the Rfam T__box (RF00230), group I self-splicing introns (RF00028), and ctRNA__pND324 (RF00238) involved in bacterial plasmid copy control all appear implausible and are also unexpected by phylogenetic criteria. Hits against Rfam small nucleolar RNAs (RF00086, RF00133, RF00309) also appeared to be false positives, as were most hits to the iron response element (RF00037) and selenocysteine insertion sequence (RF00031) families. Other families not discussed here or in Table S6 yielded no hits above threshold. See http://www.cs.washington.edu/homes/ruzzo/papers/Tthermophila for full details about the ncRNA scans. It should be noted that our annotation approach may be prone to reporting ncRNA pseudogenes and that its accuracy may be affected by the high AT content of the genome.

Protein-coding gene finding and coding region analysis. The gene finder TIGRscan ([181], since renamed GeneZilla) was trained for *T. thermophila* using a two-phase bootstrapping process [182], due to the dearth of curated training data available at the time. In the first round of training (termed "long-ORFs"), all parameters were estimated from a set of 193 full-length cDNAs from the apicomplexan *P. falciparum* (including surrounding regions from the genomic sequence; 1.6 Mb total) except for the exon state, which was trained on 2,130 nonoverlapping, long ORFs (each at least 3,000 bp in length). The default polyadenylation signal state and TATA-box state for this gene finder utilize human TRANSFAC weight matrices [183]; these were not modified. The gene finder was then used to predict genes in the raw *T. thermophila* genomic sequence, and the predictions were used to bootstrap the parameter estimation during the second round of training (termed "hybrid"). Sixty curated *T. thermophila* genes which became available during the second round of training were analyzed and their coding statistics were used to improve the exon state by averaging with the original long-ORF statistics, appropriately weighted to eliminate length bias. Exon length distributions were estimated from the 60-gene set, with appropriate smoothing. Interpolated and noninterpolated Markov chains [184] were utilized by the content states, with the order of dependency (3rd for exons and introns, 0th for intergenic, and 1st for UTR) selected so as to optimize prediction accuracy on the 60-gene set. Splice site and start/stop codon states were re-trained from pooled data consisting of the 60 curated genes and the original *P. falciparum* training data, using an 80%:20% *T. thermophila*/*P. falciparum* weighting to mitigate the effects of overtraining due to small sample sizes in the sixty gene set. Weight matrices utilized by the latter states were reduced to approximately 22 bp when it was noticed that longer matrices interfered with the prediction of short introns. The "hybrid" and "long-ORFs" parameterizations were tested on a set of 300 partial genes inferred from ESTs that were assembled against the chromosomes using the *PASA* program [52]. The "hybrid" parameterization was chosen because it was about three times more accurate at the exon level than "long-ORFs" (see Table S16).

Multivariate analysis of codon usage was performed with the

codonW package (http://codonw.sourceforge.net). Correspondence analysis of relative synonymous codon usage values was carried out to examine the major source of codon usage variation. Amino acid composition of the predicted aggregate proteome was compared with the corresponding data downloaded from dictyBase for the slime mold *D. discoideum* and from Ensembl for *Homo sapiens*.

To find candidate tandem gene duplicates, we analyzed pairwise alignments between neighboring genes using BLASTP. An all-versus-all BLASTP search was performed using all *Tetrahymena* gene-encoded proteins, requiring a maximum E-value of 1e−20, and reporting the best 20 matches. Matching genes found at adjacent genome locations were chained together and reported as candidate tandem gene arrays, allowing only a total of two nonmatching genes to intervene matching genes in a single array.

A Lek clustering algorithm [169] was applied for paralogous gene family classification of the predicted proteins in the *T. thermophila* genome. All predicted proteins were searched with BLASTP against each other. Links were established between genes at an E-value cutoff of 1×10^{-20}. Lek similarity scores, which were defined as the number of BLASTP hits shared by any pair of proteins divided by the combined number of hits for either of the two genes, were calculated for all pairs of proteins. The links for which the Lek similarity scores were above a cutoff of 0.66 were used to build gene family clusters by a single-linkage clustering algorithm. Biological function roles were assigned to the gene families based on the top BLASTP hits for individual genes in each family against a nonredundant protein database.

Organelle-derived genes and APIS. Searches for plastid and mitochondrial related genes were performed using the APIS program. APIS (J. H. Badger, unpublished data) is a system that automatically generates and summarizes phylogenetic trees for each gene in a genome. It is implemented as a series of Ruby scripts, and the results are viewable on an internal Web server which allows the user to explore the data and results in an interactive manner. APIS obtains homologs by comparing each query protein against a database of proteins from complete genomes, and extracting the full length sequences of homologs with E-values less than 1e−10. The homologs are then aligned by MUSCLE [185] and bootstrapped neighbor-joining trees are produced using QuickTree [186]. As QuickTree (unlike most programs) produces bootstrapped trees with meaningful branch lengths, the trees are then midpoint rooted. Then a taxonomic analysis is performed of the proteins that are neighbors in the tree with the query protein. This analysis makes use of the NCBI taxonomy assigned to the other proteins in the tree. For each taxonomic level (e.g., kingdom, phylum, class, etc.), the query protein is assigned to a bin. If in the tree the query protein is within a clade of sequences that are all from group X (for the taxonomic level being examined) then the query protein is placed in a bin labeled "contained within group X." If the query protein branches next to (but not within) a clade of sequences from the same group, it is placed in a bin labeled "outgroup of X." If the neighbors of the query sequence are in multiple groups, no binning is done for that taxonomic level.

Candidates for mitochondrially derived genes were separately identified by BLASTP searches using known mitochondrial proteins as queries [187,188]. Phylogenetic trees were then constructed for individual candidates in the context of all completely sequenced genomes and representatives of mitochondria. Genes whose closest neighbors were exclusively α-proteobacteria and/or mitochondria were classified as possibly mitochondrion derived.

Analysis of repetitive DNA and TEs. The location and characterization of tandem minisatellite and microsatellite repeats were done using Tandem Repeats Finder [189], using the default parameter values. The location, length, period size, %GC, and consensus sequence of each repeat were extracted for all scaffolds and listed with the scaffold number and size. Vmatch (http://www.vmatch.de) was used to search for repeats that are at least 50 bp long and 100% identical (Table S17). We note that repeats that are larger than the average insert size of our libraries would not be able to be uniquely placed into any assembly by the Celera Assembler and thus do not show up in our analysis.

The *T. thermophila* genome was searched against two sets of TEs using BLASTN and/or TBLASTN [190], with default parameters and E-value cutoff at 1×10^{-5}. One of the TE sets consisted of 12 complete or partial ciliate TEs, namely *Tec1, Tec2,* and *Tec3* from *Moneuplotes crassus, TBE1* from *O. fallax,* and *REP1, REP2.2, REP3, REP6, TIE1, TIE2, TIE3,* and *Tlr* from *T. thermophila* [90,91,191,192]. The other TE set consisted of 44 representative elements of the transposon superfamily *mariner/Tc1/IS630* [192], including members of the *mariner, Tc1,* DD39D (plant), DD37D (nematodes and insects), and DD37D (mosquitoes), Ant1/*Tec,* and *Pogo* families. In addition, the genome was scanned for homology to TE-encoded ORFs using PSI-TBLASTN [190]. Briefly, a reference ORF from each major family of autonomous transposons and retrotransposons was searched against the nonredundant protein database using BLAST-PGP with two iterations, generating a TE ORF family-specific profile. Each reference TE ORF and corresponding family profile were searched against the genomic sequence using PSI-TBLASTN, and all matches with E-value at most 1e−5 were captured for subsequent analysis. Finally, a few scaffolds with putatively complete transposases belonging to the *mariner/Tc1/IS630* superfamily were further investigated for the presence of the inverted terminal repeats (ITRs) that typically flank these elements. Identification of paired ITRs was done using Owen [193] and searches were done against known consensus ITR sequences of *mariner* and *Tc1* elements to find individual ITRs.

Analysis of functional categories with gene family expansions. Protein kinase genes were identified by comparison of peptide predictions to a set of protein kinase profile hidden Markov models [104] and by BLAST against divergent kinase sequences. A small number of gene predictions were split or fused to adjacent predictions based on presence of split or multiple kinase domains. Kinases were classified by comparison of kinase domain sequences to a set of group-, family-, and subfamily-specific hidden Markov models as well as by BLAST-based clustering of *T. thermophila* and previously classified kinases.

Predicted protein sequences were searched against a curated database of membrane transport proteins [113] for similarity to known or putative transport proteins using BLASTP. All proteins with significant hits (E-value less than 0.001) were collected and searched against the NCBI nonredundant protein and Pfam databases [194]. Transmembrane protein topology was predicted by TMHMM [195]. A Web-based interface was implemented to facilitate the annotation processes, which incorporates number of hits to the transporter database; BLAST and hidden Markov model search E-value and score; number of predicted transmembrane segments; and the description of top hits to the nonredundant protein database (http://www.membranetransport.org) [113,196].

A total of over 30,000 sequences of characterized and predicted proteases were obtained from the Merops database (http://www.merops.ac.uk, release 7.00) [119]. These sequences were searched against the *T. thermophila* predicted protein sequences using BLASTP with default settings and an E-value cutoff of less than 10^{-10} for defining protease homologs. Partial sequences (less than 80% of full-length) and redundant sequences were excluded. The domain/motif organization of predicted *T. thermophila* proteases was revealed by an InterPro search. For each putative protease, the known protease sequence or domain with the highest similarity was used as a reference for annotation; the catalytic type and protease family were predicted in accordance with the classification in Merops, and the enzyme was named in accordance with SWISS-PROT enzyme nomenclature (http://www.expasy.ch/cgi-bin/lists?peptidas.txt) and literature.

Tubulin superfamily genes were identified by a BLASTP search using *T. thermophila* α-tubulin Atu1p as the query. Twenty-one candidate predicted ORFs were identified, but two showed only moderate sequence similarity to either the amino- (TTHERM_00834920) or the carboxyl- (TTHERM_00896110) terminal halves of α- or β- tubulin and were not considered further. The 19 remaining were aligned with representative tubulins from other organisms and a neighbor-joining tree constructed using default settings of ClustalX (version 1.81) with 1,000 bootstrap runs. A prokaryotic tubulin ortholog, *Escherichia coli* FtsZ, was used as the outgroup (see Figure 7).

Using dynein subunit sequences obtained in the green alga *C. reinhardtii* or in other species when appropriate, we searched the *T. thermophila* MAC genome for orthologous sequence with TBLASTN. Candidate sequences were aligned with the sequences available in the databases of dynein subunits characterized in other experimental systems. Exon-intron borders were first approximated using the characteristics of the 64 introns previously experimentally determined in three dynein heavy chains, DYH1, DYH2, and DYH4. The 64 *T. thermophila* introns are AT rich (average 88%), are bounded by 5′-GT and AG-3′ and are relatively short (average 80 nucleotides; range, 50 to 332). The exon-intron borders and the expression of each gene were confirmed by RNA-directed PCR and, if necessary, sequencing of the amplified RT-PCR product. The verification of the exon-intron organizations of most of the heavy chains has not been completed.

Peptide sequence of Rab1A from *H. sapiens* was used to query *T. thermophila* gene predictions using BLASTP. Candidate Rab homologs were screened to include predicted proteins with complete Rab domains. These sequences were individually used in BLASTP searches

of GenBank to confirm that Rab proteins from another species were the closest match. The minimum E score cut-off was 5e−13, but the majority of homologs scored better than 1e−30. The top scoring Rab1 homolog from *T. thermophila* (TTHERM_00316280) was used in an additional BLASTP search of the *T. thermophila* genome to confirm that all Rab homologs were identified by the initial query. Homologs of other GTPases in the Rabl, Ral, Rap, Ras, Rho, and Arf families began to appear along with the lower scoring Rab homologs and were discarded from the set. Rab protein sequences from *H. sapiens* (Ensembl database), *Drosophila melanogaster* (Flybase), and *S. cerevisiae* (*Saccharomyces* Genome Database), along with those identified as described above from *T. thermophila*, were aligned using ClustalX. The alignment was refined by eye and gaps removed. The tree in Figure S7 was generated using the neighbor-joining module in Phylip 3.6. Trees constructed using maximum-likelihood and parsimony methods largely corroborated this topology. *T. thermophila* Rab homologs associated with clades of previously identified Rabs were given putative names where consistent BLASTP results were evident and are arranged in Table S15 according to functional groups. Preliminary annotations from the TGD were queried to identify predicted coat protein homologs. Others were identified in queries with peptide sequence from *D. melanogaster* homologs. *T. thermophila* homologs were used in BLASTP queries of GenBank to confirm annotations. Further analysis of AP subunits, clathrin, and dynamin-related proteins is found in [96].

Sequence availability. All of the sequences, assemblies, and gene predictions can be downloaded from the TIGR ftp site (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/t_thermophila). The sequence reads and traces can be downloaded from the NCBI trace archive at ftp://ftp.ncbi.nih.gov/pub/TraceDB/tetrahymena_thermophila. Assemblies, sequence reads, and gene predictions can be searched using multiple similarity search methods at the TIGR, TGD, and NCBI Web sites. Sequences are also available in Genbank (see below).

Supporting Information

Figure S1. Nucleotide Composition

(A) Scaffolds larger than 1 Mb were sorted by size and concatenated to make a pseudo molecule. Statistics of nucleotide composition were calculated for 2,000 bp sliding windows with a shift length of 1,000 bp. Yellow, GC skew; blue, GC%; purple, χ^2 score. The green lines delimit the scaffolds (long) or contigs within each scaffold (short).
(B) Analysis of three *T. thermophila* scaffolds of diverse size. Red boxes, genes on forward strand; green boxes, genes on reverse strand; blue, χ^2 score; orange, GC%; brown, GC skew; salmon, AT skew. The vertical light gray lines delimit contigs within each scaffold. Scaffold sizes: 8254645, 1,076 kb; 8254654, 510 kb; 8254072, 37.3 kb.

Found at DOI: 10.1371/journal.pbio.0040286.sg001 (246 KB PDF).

Figure S2. Gene Density Distribution

Using scaffolds larger than 100 kb, the percentage of predicted gene coding sequence was calculated within 10-kb windows. For the overall gene density (black bars), a sliding 10-kb window was applied at 2-kb intervals. Gray bars represent gene density in the 10-kb adjacent to each telomere.

Found at DOI: 10.1371/journal.pbio.0040286.sg002 (92 KB PDF).

Figure S3. Intron Size Distribution

Comparison of the percentage of introns in various size classes for both ab initio predicted genes (gray bars) and introns confirmed by EST sequencing (black bars).

Found at DOI: 10.1371/journal.pbio.0040286.sg003 (17 KB PDF).

Figure S4. Expression of tRNA and Other ncRNAs

(A) tRNA charging and expression. Total RNA was harvested from *T. thermophila* in log-phase growth (lanes 1 and 2) or after resuspension in 10 mM Tris starvation buffer for the times indicated. Total RNA samples were resolved by acid/urea acrylamide gel electrophoresis and transferred to nylon membrane; the same total RNA sample either untreated or deacylated at alkaline pH was used for lanes 1 and 2. Probing was performed using end-radiolabeled oligonucleotides specific for the tRNA of interest.
(B) Expression levels of ncRNAs under various conditions. Total RNA was harvested from *T. thermophila* under the growth or development conditions indicated, resolved, transferred, and probed as in (A). As an internal control for even loading, the same blot was hybridized to detect tRNA-Sec and SRP RNA (RNA PolIII transcripts found predominantly in the cytoplasm and involved in translation) and also to U1 and U2 snRNAs (RNA PolII transcripts found predominantly in the nucleus and involved in mRNA splicing).

Found at DOI: 10.1371/journal.pbio.0040286.sg004 (420 KB PDF).

Figure S5. Distribution of Repeat Content versus Scaffold Size

Orange points represent scaffolds that have been capped with telomeres at both ends.

Found at DOI: 10.1371/journal.pbio.0040286.sg005 (30 KB PDF).

Figure S6. Expansion of the Polo Kinase Family in *T. thermophila* Compared with Selected Eukaryotes

Neighbor-joining tree built from ClustalW alignment of polo kinase domains. Species abbreviations: Hs, *H. sapiens;* Dm, *D. melanogaster;* Ce, *Caenorhabditis elegans;* Sc, *S. cerevisiae;* Dd, *D. discoideum;* Tt, *T. thermophila*. Note that *T. thermophila* has multiple members of both the polo and sak subfamilies, and that even within the *T. thermophila*–specific cluster, sequences are as divergent as orthologs from vertebrates and lower metazoans. The bar indicates scale of average substitutions per site.

Found at DOI: 10.1371/journal.pbio.0040286.sg006 (71 KB PDF).

Figure S7. Phylogenetic Analysis of Rabs

Unrooted neighbor-joining tree for Rab GTPases. Bootstrap values over 40% (from 100 replicates) are indicated near corresponding branches. Predicted *T. thermophila* genes are in bold. Other Rabs are from *H. sapiens* (Hs), *D. melanogaster* (Dm), and *S. cerevisiae* (Sc). Proposed Rab families [157] are shown in colored blocks. Asterisks indicate Rabs for which there is functional evidence (**) or at least localization data (*) consistent with their groupings. *T. thermophila* genes cluster with the members of each Rab family except VII and IV (not shown in a box). There are three clades comprised exclusively of *T. thermophila* gene predictions (clades I, II, and III) shown in dark gray boxes.

Found at DOI: 10.1371/journal.pbio.0040286.sg007 (39 KB PDF).

Table S1. Genomic DNA Libraries

Found at DOI: 10.1371/journal.pbio.0040286.st001 (28 KB DOC).

Table S2. Statistics on Chromosome Assemblies and Satellite Repeats

Found at DOI: 10.1371/journal.pbio.0040286.st002 (52 KB DOC).

Table S3. Scaffolds Capped by Telomeres

Found at DOI: 10.1371/journal.pbio.0040286.st003 (352 KB DOC).

Table S4. Matches of RAPD DNA Polymorphisms to Scaffolds

Found at DOI: 10.1371/journal.pbio.0040286.st004 (167 KB DOC).

Table S5. *T. thermophila* ESTs, including Available GenBank Entries

Found at DOI: 10.1371/journal.pbio.0040286.st005 (30 KB DOC).

Table S6. ncRNAs

(A) 5S.
(B) tRNA.
(C) Other ncRNAs.
(D) tRNA gene IDs.

Found at DOI: 10.1371/journal.pbio.0040286.st006 (1.0 MB DOC).

Table S7. Genes Predicted to Be Highly Expressed on the Basis of Codon Usage Bias

Found at DOI: 10.1371/journal.pbio.0040286.st007 (388 KB DOC).

Table S8. Likely Mitochondrion-Derived Genes from the *T. thermophila* Macronuclear Genome

Found at DOI: 10.1371/journal.pbio.0040286.st008 (114 KB DOC).

Table S9. Scaffolds with Similarity to Members of the *mariner/Tc1/IS630* Superfamily

Found at DOI: 10.1371/journal.pbio.0040286.st009 (73 KB DOC).

Table S10. Recent Gene Duplications

Found at DOI: 10.1371/journal.pbio.0040286.st010 (1.9 MB DOC).

Table S11. Expanded Versions of Tables 5 through 8, including TIGR and GenBank IDs for All the Identified Genes

(A) Kinases.
(B) Membrane transporters.

(C) Proteases.
(D) Cytoskeletal related.

Found at DOI: 10.1371/journal.pbio.0040286.st011 (3.6 MB DOC).

Table S12. Human Disease Genes with Orthologs in *T. thermophila*, but Not the Yeast *S. cerevisiae*

Found at DOI: 10.1371/journal.pbio.0040286.st012 (90 KB DOC).

Table S13. Dynein Subunit Genes in *T. thermophila*

Found at DOI: 10.1371/journal.pbio.0040286.st013 (134 KB DOC).

Table S14. Membrane Traffic Component Homologs in *T. thermophila*

Found at DOI: 10.1371/journal.pbio.0040286.st014 (59 KB DOC).

Table S15. Rab Homologs in the *T. thermophila* Genome Assembly

Found at DOI: 10.1371/journal.pbio.0040286.st015 (159 KB DOC).

Table S16. Testing Different Gene Finder Parameterizations

Found at DOI: 10.1371/journal.pbio.0040286.st016 (25 KB DOC).

Table S17. The 50 Longest 100% Identical Repeats

Found at DOI: 10.1371/journal.pbio.0040286.st017 (93 KB DOC).

Accession Numbers

The GenBank (http://www.ncbi.nlm.nih.gov/Genbank) accession numbers for the *T. thermophila* genes are TTHERM_00047660, 00141160, 00279820, 00486500, 00522580, and 00823430 and for three dynein heavy chains, DYH1, DYH2, and DYH4, are AF346733, AY770505, and AF072878, respectively. The sequence contigs (AAGF01000001 to AAGF01002955), the scaffold assemblies (CH445395 to CH445797 and CH670346 to CH671913), and the gene predictions (EAR80512 to EAS07932) are available from GenBank. The Gene Identification numbers in Figure 7 obtained from JGI Chlamy v2.0 (http://genome.jgi-psf.org/chlre2/chlre2.home.html) are Ec_FtsZ, 16128088; Dm_alpha-1, 135396; Hs_alpha-1, 5174477; Cr_alpha-1, 135394; Tb_alpha,135440; Sc_alpha, 1729835; Pt_alpha, 1460090; Dm_beta-1, 158739; Hs_beta-1, 135448; Cr_beta, 8928401; Tb_beta, 135500; Pt_beta-1, 417854; Sc_beta, 1174608; Dm_gamma-1, 45644955; Hs_gamma-1, 31543831; Sc_gamma, 1729859; Cr_gamma, 8928436; Pt_delta, 10637981; Hs_delta, 50592998; Cr_delta, 75277286; Tb_delta, 13508430; Hs_epsilon, 7705915; Pt_epsilon, 18477270; Tb_epsilon, 259797; Xl_eta, 4266842; Pt_eta, 9501681; Tb_zeta, 7341314; Pt_iota, 18478276; Pt_theta, 18478274; Pt_kappa, 32812838; and Cr_epsilon (C_460065). The Ensembl Gene ID (http://www.ensembl.org) for Rab1A from *H. sapiens* is ENSG00000138069.

Acknowledgments

We would like to acknowledge the *Tetrahymena* research community and the members of our *Tetrahymena* Scientific Advisory Board for advice, support, encouragement, and assistance. In addition, we would like to specifically acknowledge many people for assistance: John Gill (sample tracking); Hean Koo (contaminant identification and trace archive and EST submission); Shannon Smith, Susan van Aken, and William Nierman (library construction); Sam Angiuoli (Web and BLAST page maintenance); Jeff Shao (database construction); Jessica Vamathevan (initial work on genome closure); Tamara Feldblyum, Terry Utterback, and the staff at the J. Craig Venter Institute's Joint Technology Center (sequencing); Lauren Smith and Jyoti Shetty (fosmid construction); Malcolm Gardner (advice); Martin Shumway (general software engineering support); Owen White (general informatics support); Leslie Bisignano and Lynn McKenna (grants support); Aimee Turner (financial operations); Tinu Akinyemi (administrative support); and Claire Fraser (for supporting the scientific research within TIGR).

Author contributions. JAE coordinated the project. JAE, RSC, EPH, and EO wrote and edited the majority of the manuscript. JAE, RSC, MW, DW, JHB, and MT performed multiple bioinformatics analyses. MT, JRW, PA, MF, RKS, and BJH coordinated the annotation. KMJ and LJT carried out genome closure. ALD and SLS generated and analyzed genome assemblies. JCS, KMK, and LS analyzed mobile DNA elements. WHM generated gene models. QR conducted analyses of membrane transporters. JMC, JG, and REP generated and analyzed ESTs. GM analyzed protein kinases. NCE and APT analyzed membrane trafficking. DJA and DEW analyzed dyneins. YW and HC analyzed proteases. KC, BAS, SRL, WLR, KW, and ZW analyzed ncRNA. DW, JG, MAG, JF, and CCT analyzed cytoskeletal associated proteins. PJK, RFW, NJP, and JHB searched for plastid-derived genes. JMC, NAS, and CJK built TGD. CdT, HFR, SCW, and RAB performed the RAPD analyses. EPH, EO, SLS, JAE, and MW examined genome structure.

Funding. This project was supported by grants to JAE from the National Science Foundation Microbial Genome Sequencing Program (EF-0240361) and the National Institutes of Health–National Institute of General Medical Sciences (R01 GM067012–03). We also acknowledge Genome Canada for support of EST library construction and sequencing through the Protist EST Project and grant RR-009231 to EO from the National Institutes of Health (the National Center for Research Resources) which supported the RAPD and Cbs work and an EO subcontract to NSF grant MCB-0132675 which supported sequence analyses related to number of chromosomes and their copy number.

Competing interests. The authors have declared that no competing interests exist.

References

1. Collins K, Gorovsky MA (2005) *Tetrahymena thermophila*. Curr Biol 15: R317–R318.
2. Nanney DL, Simon EM (2000) Laboratory and evolutionary history of *Tetrahymena thermophila*. Methods Cell Biol 62: 3–25.
3. Zaug AJ, Cech TR (1986) The intervening sequence RNA of *Tetrahymena* is an enzyme. Science 231: 470–475.
4. Blackburn EH, Gall JG (1978) A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in *Tetrahymena*. J Mol Biol 120: 33–53.
5. Yao MC, Yao CH (1981) Repeated hexanucleotide C-C-C-C-A-A is present near free ends of macronuclear DNA of *Tetrahymena*. Proc Natl Acad Sci U S A 78: 7436–7439.
6. Greider CW, Blackburn EH (1985) Identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts. Cell 43: 405–413.
7. Brownell JE, Zhou J, Ranalli T, Kobayashi R, Edmondson DG, et al. (1996) *Tetrahymena* histone acetyltransferase A: A homolog to yeast Gcn5p linking histone acetylation to gene activation. Cell 84: 843–851.
8. Asai DJ, Forney JD, editors (2000) *Tetrahymena thermophila*. San Diego: Academic Press. 580 p.
9. Turkewitz AP, Orias E, Kapler G (2002) Functional genomics: The coming of age for *Tetrahymena thermophila*. Trends Genet 18: 35–40.
10. Kim K, Weiss LM (2004) *Toxoplasma gondii*: The model apicomplexan. Int J Parasitol 34: 423–432.
11. Donald RG, Roos DS (1998) Gene knock-outs and allelic replacements in *Toxoplasma gondii*: HXGPRT as a selectable marker for hit-and-run mutagenesis. Mol Biochem Parasitol 91: 295–305.
12. Radke JR, Behnke MS, Mackey AJ, Radke JB, Roos DS, et al. (2005) The transcriptome of *Toxoplasma gondii*. BMC Biol 3: 26.
13. Peterson DS, Gao Y, Asokan K, Gaertig J (2002) The circumsporozoite protein of *Plasmodium falciparum* is expressed and localized to the cell surface in the free-living ciliate *Tetrahymena thermophila*. Mol Biochem Parasitol 122: 119–126.
14. Prescott DM (1994) The DNA of ciliated protozoa. Microbiol Rev 58: 233–267.
15. Martindale DW, Allis CD, Bruns PJ (1982) Conjugation in *Tetrahymena thermophila*. A temporal analysis of cytological stages. Exp Cell Res 140: 227–236.
16. Yao MC, Chao JL (2005) RNA-guided DNA deletion in *Tetrahymena*: An RNAi-based mechanism for programmed genome rearrangements. Annu Rev Genet 39: 537–559.
17. Yao MC, Duharcourt S, Chalker DL (2002) Genome-wide rearrangements of DNA in ciliates. In: Craig N, Craigie R, Gellert M, Lambowitz A, editors. Mobile DNA II. Herndon (Virginia): ASM Press. pp. 730–758.
18. Yao MC, Choi J, Yokoyama S, Austerberry CF, Yao CH (1984) DNA elimination in *Tetrahymena*: A developmental process involving extensive breakage and rejoining of DNA at defined sites. Cell 36: 433–440.
19. Yao MC, Gorovsky MA (1974) Comparison of the sequences of macro- and micronuclear DNA of *Tetrahymena* pyriformis. Chromosoma 48: 1–18.
20. Iwamura Y, Sakai M, Muramatsu M (1982) Rearrangement of repeated DNA sequences during development of macronucleus in *Tetrahymena thermophila*. Nucleic Acids Res 10: 4279–4291.
21. Jenuwein T (2002) Molecular biology. An RNA-guided pathway for the epigenome. Science 297: 2215–2218.
22. Selker EU (2003) Molecular biology. A self-help guide for a trim genome. Science 300: 1517–1518.
23. Fan Q, Yao MC (2000) A long stringent sequence signal for programmed chromosome breakage in *Tetrahymena thermophila*. Nucleic Acids Res 28: 895–900.
24. Hamilton EP, Williamson S, Dunn S, Merriam V, Lin C, et al. (2006) The highly conserved family of *Tetrahymena thermophila* chromosome breakage

elements contains an invariant 10-base-pair core. Eukaryot Cell 5: 771–780.
25. Yao MC, Yao CH, Monks B (1990) The controlling sequence for site-specific chromosome breakage in *Tetrahymena*. Cell 63: 763–772.
26. Fan Q, Yao M (1996) New telomere formation coupled with site-specific chromosome breakage in *Tetrahymena thermophila*. Mol Cell Biol 16: 1267–1274.
27. Yu GL, Blackburn EH (1991) Developmentally programmed healing of chromosomes by telomerase in *Tetrahymena*. Cell 67: 823–832.
28. Altschuler MI, Yao MC (1985) Macronuclear DNA of *Tetrahymena thermophila* exists as defined subchromosomal-sized molecules. Nucleic Acids Res 13: 5817–5831.
29. Conover RK, Brunk CF (1986) Macronuclear DNA molecules of *Tetrahymena thermophila*. Mol Cell Biol 6: 900–905.
30. Kapler GM (1993) Developmentally regulated processing and replication of the *Tetrahymena* rDNA minichromosome. Curr Opin Genet Dev 3: 730–735.
31. Doerder FP, Deak JC, Lief JH (1992) Rate of phenotypic assortment in *Tetrahymena thermophila*. Dev Genet 13: 126–132.
32. Ray C Jr (1956) Preparation of chromosomes of *Tetrahymena* pyriformis for photomicrography. Stain Technol 31: 271–274.
33. LaFountain JR Jr, Davidson LA (1979) An analysis of spindle ultrastructure during prometaphase and metaphase of micronuclear division in *Tetrahymena*. Chromosoma 75: 293–308.
34. LaFountain JR Jr, Davidson LA (1980) An analysis of spindle ultrastructure during anaphase of micronuclear division in *Tetrahymena*. Cell Motil 1: 41–61.
35. Mochizuki K, Gorovsky MA (2004) Small RNAs in genome rearrangement in *Tetrahymena*. Curr Opin Genet Dev 14: 181–187.
36. Orias E (2000) Toward sequencing the *Tetrahymena* genome: Exploiting the gift of nuclear dimorphism. J Eukaryot Microbiol 47: 328–333.
37. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, et al. (2000) A whole-genome assembly of *Drosophila*. Science 287: 2196–2204.
38. Brunk CF, Lee LC, Tran AB, Li J (2003) Complete sequence of the mitochondrial genome of *Tetrahymena thermophila* and comparative methods for identifying highly divergent genes. Nucleic Acids Res 31: 1673–1682.
39. Engberg J, Nielsen H (1990) Complete sequence of the extrachromosomal rDNA molecule from the ciliate *Tetrahymena thermophila* strain B1868VII. Nucleic Acids Res 18: 6915–6919.
40. Wong L, Klionsky L, Wickert S, Merriam V, Orias E, et al. (2000) Autonomously replicating macronuclear DNA pieces are the physical basis of genetic coassortment groups in *Tetrahymena thermophila*. Genetics 155: 1119–1125.
41. Cassidy-Hanley D, Bisharyan Y, Fridman V, Gerber J, Lin C, et al. (2005) Genome-wide characterization of *Tetrahymena thermophila* chromosome breakage sites. II. Physical and genetic mapping. Genetics 170: 1623–1631.
42. Yao MC, Zheng K, Yao CH (1987) A conserved nucleotide sequence at the sites of developmentally regulated chromosomal breakage in *Tetrahymena*. Cell 48: 779–788.
43. Karrer KM (2000) *Tetrahymena* genetics: Two nuclei are better than one. Methods Cell Biol 62: 127–186.
44. Cervantes MD, Xi X, Vermaak D, Yao MC, Malik HS (2006) The CNA1 histone of the ciliate *Tetrahymena thermophila* is essential for chromosome segregation in the germline micronucleus. Mol Biol Cell 17: 485–497.
45. Pryde FE, Gorham HC, Louis EJ (1997) Chromosome ends: All the same under their caps. Curr Opin Genet Dev 7: 822–828.
46. Wellinger RJ, Sen D (1997) The DNA structures at the ends of eukaryotic chromosomes. Eur J Cancer 33: 735–749.
47. Barry JD, Ginger ML, Burton P, McCulloch R (2003) Why are parasite contingency genes often associated with telomeres? Int J Parasitol 33: 29–45.
48. Gao W, Khang CH, Park SY, Lee YH, Kang S (2002) Evolution and organization of a highly dynamic, subtelomeric helicase gene family in the rice blast fungus *Magnaporthe grisea*. Genetics 162: 103–112.
49. Mefford HC, Trask BJ (2002) The complex structure and dynamic evolution of human subtelomeres. Nat Rev Genet 3: 91–102.
50. Teunissen AW, Steensma HY (1995) Review: The dominant flocculation genes of *Saccharomyces cerevisiae* constitute a new subtelomeric gene family. Yeast 11: 1001–1013.
51. Louis EJ (1995) The chromosome ends of *Saccharomyces cerevisiae*. Yeast 11: 1553–1573.
52. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, et al. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31: 5654–5666.
53. Calzone FJ, Stathopoulos VA, Grass D, Gorovsky MA, Angerer RC (1983) Regulation of protein synthesis in *Tetrahymena*. RNA sequence sets of growing and starved cells. J Biol Chem 258: 6899–6905.
54. Zagulski M, Nowak JK, Le Mouel A, Nowacki M, Migdalski A, et al. (2004) High coding density on the largest *Paramecium tetraurelia* somatic chromosome. Curr Biol 14: 1397–1404.
55. Erdmann VA, Wolters J, Huysmans E, Vandenberghe A, De Wachter R (1984) Collection of published 5S and 5.8S ribosomal RNA sequences. Nucleic Acids Res 12: r133–r166.
56. Luehrsen KR, Fox GE, Woese CR (1980) The sequence of *Tetrahymena thermophila* 5S ribosomal ribonucleic acid. Curr Microbiol 4: 123–126.
57. Kimmel AR, Gorovsky MA (1976) Numbers of 5S and tRNA genes in macro- and micronuclei of *Tetrahymena pyriformis*. Chromosoma 54: 327–337.
58. Horowitz S, Gorovsky MA (1985) An unusual genetic code in nuclear genes of *Tetrahymena*. Proc Natl Acad Sci U S A 82: 2452–2455.
59. Driscoll DM, Copeland PR (2003) Mechanism and regulation of selenoprotein synthesis. Annu Rev Nutr 23: 17–40.
60. Hatfield DL, Gladyshev VN (2002) How selenium has altered our understanding of the genetic code. Mol Cell Biol 22: 3565–3576.
61. Shrimali RK, Lobanov AV, Xu XM, Rao M, Carlson BA, et al. (2005) Selenocysteine tRNA identification in the model organisms *Dictyostelium discoideum* and *Tetrahymena thermophila*. Biochem Biophys Res Commun 329: 147–151.
62. Wuitschick JD, Karrer KM (1999) Analysis of genomic G + C content, codon usage, initiator codon context and translation termination sites in *Tetrahymena thermophila*. J Eukaryot Microbiol 46: 239–247.
63. Wuitschick JD, Karrer KM (2000) Codon usage in *Tetrahymena thermophila*. Methods Cell Biol 62: 565–568.
64. Eichinger L, Pachebat JA, Glockner G, Rajandream MA, Sucgang R, et al. (2005) The genome of the social amoeba *Dictyostelium discoideum*. Nature 435: 43–57.
65. Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T (2001) Codon usage and tRNA genes in eukaryotes: Correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J Mol Evol 53: 290–298.
66. Lynch M, Conery JS (2003) The origins of genome complexity. Science 302: 1401–1404.
67. Katz LA, Snoeyenbos-West O, Doerder FP (2006) Patterns of protein evolution in *Tetrahymena thermophila*: Implications for estimates of effective population size. Mol Biol Evol 23: 608–614.
68. Fast NM, Xue L, Bingham S, Keeling PJ (2002) Re-examining alveolate evolution using multiple protein molecular phylogenies. J Eukaryot Microbiol 49: 30–37.
69. Gajadhar AA, Marquardt WC, Hall R, Gunderson J, Ariztia-Carmona EV, et al. (1991) Ribosomal RNA sequences of *Sarcocystis muris*, *Theileria annulata* and *Crypthecodinium cohnii* reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates. Mol Biochem Parasitol 45: 147–154.
70. Gardner MJ, Williamson DH, Wilson RJ (1991) A circular DNA in malaria parasites encodes an RNA polymerase like that of prokaryotes and chloroplasts. Mol Biochem Parasitol 44: 115–123.
71. Cavalier-Smith T (1999) Principles of protein and lipid targeting in secondary symbiogenesis: Euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J Eukaryot Microbiol 46: 347–366.
72. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419: 498–511.
73. Regoes A, Zourmpanou D, Leon-Avila G, van der Giezen M, Tovar J, et al. (2005) Protein import, replication, and inheritance of a vestigial mitochondrion. J Biol Chem 280: 30557–30563.
74. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815.
75. Dyall SD, Brown MT, Johnson PJ (2004) Ancient invasions: From endosymbionts to organelles. Science 304: 253–257.
76. Ralph SA, van Dooren GG, Waller RF, Crawford MJ, Fraunholz MJ, et al. (2004) Tropical infectious diseases: Metabolic maps and functions of the *Plasmodium falciparum* apicoplast. Nat Rev Microbiol 2: 203–216.
77. Erwin JA, Beach D, Holz GG Jr (1966) Effect of dietary cholesterol on unsaturated fatty acid biosynthesis in a ciliated protozoan. Biochim Biophys Acta 125: 614–616.
78. Holz GG Jr, Erwin J, Rosenbaum N, Aaronson S (1962) Triparanol inhibition of *Tetrahymena*, and its prevention by lipids. Arch Biochem Biophys 98: 312–322.
79. Holz GG Jr, Wagner B, Erwin J, Britt JJ, Bloch K (1961) Sterol requirements of a ciliate *Tetrahymena corlissi* Th-X. I. A nutritional analysis of the sterol requirements of *T. corlissi* Th-X. II. Metabolism of tritiated lopohenol in *T. corlissi* Th-X. Comp Biochem Physiol 2: 202–217.
80. Corliss JO (1979) The impact of electron microscopy on ciliate systematics. Am Zool 19: 573–587.
81. Lynn DH (1981) The organization and evolution of microtubular organelles in ciliated protozoa. Biol Rev 56: 243–292.
82. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, et al. (2004) Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. Science 304: 441–445.
83. Huang J, Mullapudi N, Lancto CA, Scott M, Abrahamsen MS, et al. (2004) Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. Genome Biol 5: R88.
84. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, et al. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422: 859–868.
85. Galagan JE, Selker EU (2004) RIP: The evolutionary cost of genome defense. Trends Genet 20: 417–423.
86. Liu Y, Song X, Gorovsky MA, Karrer KM (2005) Elimination of foreign

DNA during somatic differentiation in *Tetrahymena thermophila* shows position effect and is dosage dependent. Eukaryot Cell 4: 421–431.
87. Mochizuki K, Fine NA, Fujisawa T, Gorovsky MA (2002) Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. Cell 110: 689–699.
88. Yao MC, Fuller P, Xi X (2003) Programmed DNA deletion as an RNA-guided system of genome defense. Science 300: 1581–1584.
89. Doerder FP, Gates MA, Eberhardt FP, Arslanyolu M (1995) High frequency of sex and equal frequencies of mating types in natural populations of the ciliate *Tetrahymena thermophila*. Proc Natl Acad Sci U S A 92: 8715–8718.
90. Fillingham JS, Thing TA, Vythilingum N, Keuroghlian A, Bruno D, et al. (2004) A nonlong terminal repeat retrotransposon family is restricted to the germ line micronucleus of the ciliated protozoan *Tetrahymena thermophila*. Eukaryot Cell 3: 157–169.
91. Wuitschick JD, Gershan JA, Lochowicz AJ, Li S, Karrer KM (2002) A novel family of mobile genetic elements is limited to the germline genome in *Tetrahymena thermophila*. Nucleic Acids Res 30: 2524–2537.
92. Pritham EJ, Feschotte C, Wessler SR (2005) Unexpected diversity and differential success of DNA transposons in four species of entamoeba protozoans. Mol Biol Evol 22: 1751–1763.
93. Silva JC, Bastida F, Bidwell SL, Johnson PJ, Carlton JM (2005) A potentially functional mariner transposable element in the protist *Trichomonas vaginalis*. Mol Biol Evol 22: 126–134.
94. Foss EJ, Garrett PW, Kinsey JA, Selker EU (1991) Specificity of repeat-induced point mutation (RIP) in *Neurospora*: Sensitivity of non*Neurospora* sequences, a natural diverged tandem duplication, and unique DNA adjacent to a duplicated region. Genetics 127: 711–717.
95. Bowman GR, Smith DG, Michael Siu KW, Pearlman RE, Turkewitz AP (2005) Genomic and proteomic evidence for a second family of dense core granule cargo proteins in *Tetrahymena thermophila*. J Eukaryot Microbiol 52: 291–297.
96. Elde NC, Morgan G, Winey M, Sperling L, Turkewitz AP (2005) Elucidation of clathrin-mediated endocytosis in *Tetrahymena* reveals an evolutionarily convergent recruitment of dynamin. PLoS Genetics 1: e52. DOI: 10.1371/journal.pgen.0010052
97. Herrmann L, Erkelenz M, Aldag I, Tiedtke A, Hartmann MW (2006) Biochemical and molecular characterisation of *Tetrahymena thermophila* extracellular cysteine proteases. BMC Microbiol 6: 19.
98. Kuribara S, Kato M, Kato-Minoura T, Numata O (2006) Identification of a novel actin-related protein in *Tetrahymena* cilia. Cell Motil Cytoskeleton 63: 437–446.
99. Lee SR, Collins K (2006) Two classes of endogenous small RNAs in *Tetrahymena thermophila*. Genes Dev 20: 28–33.
100. Stemm-Wolf AJ, Morgan G, Giddings TH Jr, White EA, Marchione R, et al. (2005) Basal body duplication and maintenance require one member of the *Tetrahymena thermophila* centrin gene family. Mol Biol Cell 16: 3606–3619.
101. Wickstead B, Gull K (2006) A "holistic" kinesin phylogeny reveals new kinesin families and predicts protein functions. Mol Biol Cell 17: 1734–1743.
102. Williams SA, Gavin RH (2005) Myosin genes in *Tetrahymena*. Cell Motil Cytoskeleton 61: 237–243.
103. Wloga D, Camba A, Rogowski K, Manning G, Jerka-Dziadosz M, et al. (2006) Members of the Nima-related kinase family promote disassembly of cilia by multiple mechanisms. Mol Biol Cell 17: 2799–2810.
104. Global analysis of protein kinase genes in sequenced genomes. Available: http://kinase.com. Accessed 15 July 2006.
105. Manning G, Plowman GD, Hunter T, Sudarsanam S (2002) Evolution of protein kinase signaling from yeast to man. Trends Biochem Sci 27: 514–520.
106. Goldberg JM, Manning G, Liu A, Fey P, Pilcher KE, et al. (2006) The dictyostelium kinome—Analysis of the protein kinases from a simple model organism. PLoS Genet 2: e38. DOI: 10.1371/journal.pgen.0020038
107. Christensen ST, Guerra CF, Awan A, Wheatley DN, Satir P (2003) Insulin receptor-like proteins in *Tetrahymena thermophila* ciliary membranes. Curr Biol 13: R50–R52.
108. Manning G, Caenepeel S (2005) Protein kinases in human disease. 2005–06 Catalog and technical reference. Beverly (Massachusetts): Cell Signaling Technologies. pp. 402–409.
109. O'Connell MJ, Krien MJ, Hunter T (2003) Never say never. The NIMA-related protein kinases in mitotic control. Trends Cell Biol 13: 221–228.
110. Okazaki N, Yan J, Yuasa S, Ueno T, Kominami E, et al. (2000) Interaction of the Unc-51-like kinase and microtubule-associated protein light chain 3 related proteins in the brain: Possible role of vesicular transport in axonal elongation. Brain Res Mol Brain Res 85: 1–12.
111. Wolanin PM, Thomason PA, Stock JB (2002) Histidine protein kinases: Key signal transducers outside the animal kingdom. Genome Biol 3: reviews3013.
112. Hanks SK (2003) Genomic analysis of the eukaryotic protein kinase superfamily: A perspective. Genome Biol 4: 111.
113. Ren Q, Kang KH, Paulsen IT (2004) TransportDB: A relational database of cellular membrane transport systems. Nucleic Acids Res 32: D284–D288.
114. Haynes WJ, Ling KY, Saimi Y, Kung C (2003) PAK paradox: *Paramecium* appears to have more K(+)-channel genes than humans. Eukaryot Cell 2: 737–745.
115. Kung C, Saimi Y (1982) The physiological basis of taxes in *Paramecium*. Annu Rev Physiol 44: 519–534.
116. Hennessey T, Machemer H, Nelson DL (1985) Injected cyclic AMP increases ciliary beat frequency in conjunction with membrane hyperpolarization. Eur J Cell Biol 36: 153–156.
117. Weber JH, Vishnyakov A, Hambach K, Schultz A, Schultz JE, et al. (2004) Adenylyl cyclases from *Plasmodium, Paramecium* and *Tetrahymena* are novel ion channel/enzyme fusion proteins. Cell Signal 16: 115–125.
118. Puente XS, Sanchez LM, Overall CM, Lopez-Otin C (2003) Human and mouse proteases: A comparative genomic approach. Nat Rev Genet 4: 544–558.
119. Rawlings ND, Tolle DP, Barrett AJ (2004) MEROPS: The peptidase database. Nucleic Acids Res 32: D160–D164.
120. Southan C (2001) A genomic perspective on human proteases. FEBS Lett 498: 214–218.
121. Barrett AJ, Rawlings ND, Woessner JF, editors (1998) Handbook of proteolytic enzymes. San Diego: Academic Press. 1666 p.
122. Wu Y, Wang X, Liu X, Wang Y (2003) Data-mining approaches reveal hidden families of proteases in the genome of malaria parasite. Genome Res 13: 601–616.
123. Bochtler M, Ditzel L, Groll M, Hartmann C, Huber R (1999) The proteasome. Annu Rev Biophys Biomol Struct 28: 295–317.
124. Gruszynski AE, DeMaster A, Hooper NM, Bangs JD (2003) Surface coat remodeling during differentiation of *Trypanosoma brucei*. J Biol Chem 278: 24665–24672.
125. LaCount DJ, Gruszynski AE, Grandgenett PM, Bangs JD, Donelson JE (2003) Expression and function of the *Trypanosoma brucei* major surface protease (GP63) genes. J Biol Chem 278: 24658–24664.
126. Yao C, Donelson JE, Wilson ME (2003) The major surface protease (MSP or GP63) of *Leishmania* sp. Biosynthesis, regulation of expression, and function. Mol Biochem Parasitol 132: 1–16.
127. Madeo F, Herker E, Maldener C, Wissing S, Lachelt S, et al. (2002) A caspase-related protease regulates apoptosis in yeast. Mol Cell 9: 911–917.
128. Frankel J (2000) Cell biology of *Tetrahymena thermophila*. Methods Cell Biol 62: 27–125.
129. Williams NE (2000) Preparation of cytoskeletal fractions from *Tetrahymena thermophila*. Methods Cell Biol 62: 441–447.
130. Dutcher SK (2003) Long-lost relatives reappear: Identification of new members of the tubulin superfamily. Curr Opin Microbiol 6: 634–640.
131. Gaertig J, Thatcher TH, McGrath KE, Callahan RC, Gorovsky MA (1993) Perspectives on tubulin isotype function and evolution based on the observation that *Tetrahymena thermophila* microtubules contain a single alpha- and beta-tubulin. Cell Motil Cytoskeleton 25: 243–253.
132. McGrath KE, Yu SM, Heruth DP, Kelly AA, Gorovsky MA (1994) Regulation and evolution of the single alpha-tubulin gene of the ciliate *Tetrahymena thermophila*. Cell Motil Cytoskeleton 27: 272–283.
133. Shang Y, Li B, Gorovsky MA (2002) *Tetrahymena thermophila* contains a conventional gamma-tubulin that is differentially required for the maintenance of different microtubule-organizing centers. J Cell Biol 158: 1195–1206.
134. Dupuis-Williams P, Fleury-Aubusson A, de Loubresse NG, Geoffroy H, Vayssie L, et al. (2002) Functional role of epsilon-tubulin in the assembly of the centriolar microtubule scaffold. J Cell Biol 158: 1183–1193.
135. Ruiz F, Dupuis-Williams P, Klotz C, Forquignon F, Bergdoll M, et al. (2004) Genetic evidence for interaction between eta- and beta-tubulins. Eukaryot Cell 3: 212–220.
136. Ruiz F, Krzywicka A, Klotz C, Keller A, Cohen J, et al. (2000) The SM19 gene, required for duplication of basal bodies in *Paramecium*, encodes a novel tubulin, eta-tubulin. Curr Biol 10: 1451–1454.
137. Duan J, Gorovsky MA (2002) Both carboxy-terminal tails of alpha- and beta-tubulin are essential, but either one will suffice. Curr Biol 12: 313–316.
138. Thazhath R, Liu C, Gaertig J (2002) Polyglycylation domain of beta-tubulin maintains axonemal architecture and affects cytokinesis in *Tetrahymena*. Nat Cell Biol 4: 256–259.
139. Xia L, Hai B, Gao Y, Burnette D, Thazhath R, et al. (2000) Polyglycylation of tubulin is essential and affects cell motility and division in *Tetrahymena thermophila*. J Cell Biol 149: 1097–1106.
140. Gibbons IR, Rowe AJ (1965) Dynein: A protein with adenosine triphosphatase activity from cilia. Science 149: 424–426.
141. Gibbons IR, Lee-Eiford A, Mocz G, Phillipson CA, Tang WJ, et al. (1987) Photosensitized cleavage of dynein heavy chains. Cleavage at the "V1 site" by irradiation at 365 nm in the presence of ATP and vanadate. J Biol Chem 262: 2780–2786.
142. King SM (2000) The dynein microtubule motor. Biochim Biophys Acta 1496: 60–75.
143. Sakato M, King SM (2004) Design and regulation of the AAA+ microtubule motor dynein. J Struct Biol 146: 58–71.
144. Asai DJ, Koonce MP (2001) The dynein heavy chain: Structure, mechanics and evolution. Trends Cell Biol 11: 196–202.
145. Asai DJ, Wilkes DE (2004) The dynein heavy chain family. J Eukaryot Microbiol 51: 23–29.
146. Sailaja G, Lincoln LM, Chen J, Asai DJ (2001) Evaluating the dynein heavy chain gene family in *Tetrahymena*. Methods Mol Biol 161: 17–27.
147. Xu W, Royalty MP, Zimmerman JR, Angus SP, Pennock DG (1999) The

dynein heavy chain gene family in *Tetrahymena thermophila*. J Eukaryot Microbiol 46: 606–611.
148. Foth BJ, Goedecke MC, Soldati D (2006) New insights into myosin evolution and classification. Proc Natl Acad Sci U S A 103: 3681–3686.
149. Janke C, Rogowski K, Wloga D, Regnard C, Kajava AV, et al. (2005) Tubulin polyglutamylase enzymes are members of the TTL domain protein family. Science 308: 1758–1762.
150. Osmani SA, Engle DB, Doonan JH, Morris NR (1988) Spindle formation and chromatin condensation in cells blocked at interphase by mutation of a negative cell cycle control gene. Cell 52: 241–251.
151. Fry AM, Meraldi P, Nigg EA (1998) A centrosomal function for the human Nek2 protein kinase, a member of the NIMA family of cell cycle regulators. EMBO J 17: 470–481.
152. Mahjoub MR, Montpetit B, Zhao L, Finst RJ, Goh B, et al. (2002) The FA2 gene of *Chlamydomonas* encodes a NIMA family kinase with roles in cell cycle progression and microtubule severing during deflagellation. J Cell Sci 115: 1759–1768.
153. Turkewitz AP (2004) Out with a bang! *Tetrahymena* as a model system to study secretory granule biogenesis. Traffic 5: 63–68.
154. Bock JB, Matern HT, Peden AA, Scheller RH (2001) A genomic perspective on membrane compartment organization. Nature 409: 839–841.
155. Ackers JP, Dhir V, Field MC (2005) A bioinformatic analysis of the RAB genes of *Trypanosoma brucei*. Mol Biochem Parasitol 141: 89–97.
156. Stenmark H, Olkkonen VM (2001) The Rab GTPase family. Genome Biol 2: reviews3007.
157. Pereira-Leal JB, Seabra MC (2001) Evolution of the Rab family of small GTP-binding proteins. J Mol Biol 313: 889–901.
158. Saito-Nakano Y, Loftus BJ, Hall N, Nozaki T (2005) The diversity of Rab GTPases in *Entamoeba histolytica*. Exp Parasitol 110: 244–252.
159. Lal K, Field MC, Carlton JM, Warwicker J, Hirt RP (2005) Identification of a very large Rab GTPase family in the parasitic protozoan *Trichomonas vaginalis*. Mol Biochem Parasitol 143: 226–235.
160. Muller HM, Kenny EE, Sternberg PW (2004) Textpresso: An ontology-based information retrieval and extraction system for biological literature. PLoS Biol 2: e309.
161. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, et al. (2002) The generic genome browser: A building block for a model organism system database. Genome Res 12: 1599–1610.
162. Dear PH, Cook PR (1993) Happy mapping: Linkage mapping using a physical analogue of meiosis. Nucleic Acids Res 21: 13–20.
163. Elliott AM, Gruchy DF (1952) The occurence of mating types in *Tetrahymena*. Biol Bull (Woods Hole, MA) 105: 301.
164. Mayo KA, Orias E (1981) Further evidence for lack of gene expression in the *Tetrahymena* micronucleus. Genetics 98: 747–762.
165. Allen SL, Gibson I (1973) Genetics of *Tetrahymena*. In: Elliott AM, editor. Biology of Tetrahymena. Stroudsburg (Pennsylvania): Dowden, Hutchinson and Ross. pp. 307–373.
166. Allen SL (1967) Genomic exclusion: A rapid means for inducing homozygous diploid lines in *Tetrahymena pyriformis*, syngen 1. Science 155: 575–577.
167. Ward N, Eisen J, Fraser C, Stackebrandt E (2001) Sequenced strains must be saved from extinction. Nature 414: 148.
168. Gorovsky MA, Yao MC, Keevert JB, Pleger GL (1975) Isolation of micro- and macronuclei of *Tetrahymena* pyriformis. Methods Cell Biol 9: 311–327.
169. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.
170. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. (2004) Versatile and open software for comparing large genomes. Genome Biol 5: R12.
171. Lynch TJ, Brickner J, Nakano KJ, Orias E (1995) Genetic map of randomly amplified DNA polymorphisms closely linked to the mating type locus of *Tetrahymena thermophila*. Genetics 141: 1315–1325.
172. Hamilton E, Bruns P, Lin C, Merriam V, Orias E, et al. (2005) Genome-wide characterization of *Tetrahymena thermophila* chromosome breakage sites. I. Cloning and identification of functional sites. Genetics 170: 1611–1621.
173. Birren B, Lai E (1993) Pulsed field gel electrophoresis—A practical guide. New York: Academic Press.
174. Eddy SR, Durbin R (1994) RNA sequence analysis using covariance models. Nucleic Acids Res 22: 2079–2088.
175. Lowe TM, Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955–964.
176. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: An RNA family database. Nucleic Acids Res 31: 439–441.
177. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, et al. (2005) Rfam: Annotating noncoding RNAs in complete genomes. Nucleic Acids Res 33: D121–D124.
178. Weinberg Z, Ruzzo WL (2004) Exploiting conserved structure for faster annotation of noncoding RNAs without loss of accuracy. Bioinformatics 20: I334–I341.
179. Orum H, Nielsen H, Engberg J (1993) Sequence and proposed secondary structure of the *Tetrahymena thermophila* U3-snRNA. Nucleic Acids Res 21: 2511.
180. Weinberg Z, Ruzzo WL (2006) Sequence-based heuristics for faster annotation of noncoding RNA families. Bioinformatics 22: 35–39.
181. Majoros WH, Pertea M, Salzberg SL (2004) TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. Bioinformatics 20: 2878–2879.
182. Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5: 59.
183. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, et al. (2003) TRANSFAC: Transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374–378.
184. Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H (1999) Interpolated Markov models for eukaryotic gene finding. Genomics 59: 24–31.
185. Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
186. Howe K, Bateman A, Durbin R (2002) QuickTree: Building huge Neighbour-Joining trees of protein sequences. Bioinformatics 18: 1546–1547.
187. Scharfe C, Zaccaria P, Hoertnagel K, Jaksch M, Klopstock T, et al. (2000) MITOP, the mitochondrial proteome database: 2000 Update. Nucleic Acids Res 28: 155–158.
188. Scharfe C, Zaccaria P, Hoertnagel K, Jaksch M, Klopstock T, et al. (1999) MITOP: Database for mitochondria-related proteins, genes and diseases. Nucleic Acids Res 27: 153–155.
189. Benson G (1999) Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res 27: 573–580.
190. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
191. Gershan JA, Karrer KM (2000) A family of developmentally excised DNA elements in *Tetrahymena* is under selective pressure to maintain an open reading frame encoding an integrase-like protein. Nucleic Acids Res 28: 4105–4112.
192. Shao H, Tu Z (2001) Expanding the diversity of the IS630-Tc1-mariner superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. Genetics 159: 1103–1115.
193. Ogurtsov AY, Roytberg MA, Shabalina SA, Kondrashov AS (2002) OWEN: Aligning long collinear regions of genomes. Bioinformatics 18: 1703–1704.
194. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: Multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 26: 320–322.
195. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J Mol Biol 305: 567–580.
196. Ren Q, Paulsen IT (2005) Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes. PLoS Comput Biol 1: e27. DOI: 10.1371/journal.pcbi.0010027
197. Baldauf SL (2003) The deep roots of eukaryotes. Science 300: 1703–1706.