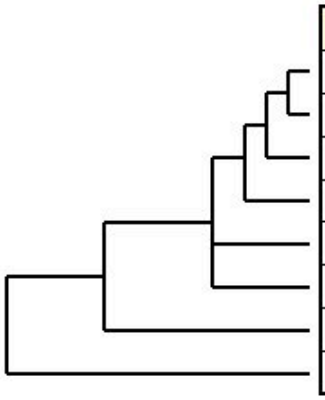
Comparative genomics of *Oxytricha* and related spirotrichous ciliates:

Minimal eukaryotic genome architectures

Thomas Doak, Yi Zhou, Estienne Swart, Laura Landweber (Princeton University)
Sean Eddy (HHMI Janelia Farm Research Campus)
Larry Klobutcher (University of Connecticut Health Center)
Rick Wilson, Elaine Mardis, Vincent Magrini (Washington University Genome Sequencing Center)

Summary

To augment the study of the extremely fragmented genomes of spirotrichous ciliates and to improve our understanding and annotation of the *Oxytricha trifallax* strain JRB310 genome, we propose the sequencing of a set of additional genomes, to various depths. This work would be based on the already supported and ongoing sequencing of the 50 Mb *Oxytricha trifallax* JRB310 macronuclear and (~50 Mb sample of the 1Gb) micronuclear genome, being performed at the Washington University Genome Sequencing Center (PI Elaine Mardis).



	MAC (~50Mb)	MIC (1Gb)	cDNA
***Oxytricha trifallax* 310**	**Currently 5X**	**2X (1)**	**in prog.**
***Oxytricha trifallax* 510 (1~10%)**	**2X (3[a])**	**-**	**in prog.**
***Oxytricha trifallax* JRB317 (1~10%)**	**2X (3[b])**	**-**	**-**
***Oxytricha trifallax* JRA55 (1~10%)**	**2X (3[b])**	**-**	**-**
***Oxytricha bifallax* (3~15%)**	**2X (3)**	**-**	**-**
***Oxytricha fallax* (3~15%)**	**2X (3)**	**-**	**25,000 (3)**
***Stylonychia lemnae* (45~75%)**	**5X (2)**	**-**	**25,000 (2)**
***Euplotes crassus* (>75%)**	**5X (4)**	**-**	**25,000 (4)**

Table 1. Requested WGS sequencing coverage with priorities (1-4) and expected neutral distances

Notes: **Priority 3** uses 454 or similar technology.
Priorities 2-4 each include 25K cDNAs for 1 genome.
a: higher status within priority 3 for shadowing
b: or a replacement at comparable distance.

We suggest (**1**) as a complement to the current sequencing of *O. trifallax*, deeper coverage of its 1 Gb germline micronuclear genome, to a full 2X coverage, rather than the mere 50 Mb of sequence that was originally approved as a companion to the *O. trifallax* macronuclear genome in March 2002 (5 years ago). At least this level of resolution will be required to permit reasonable accuracy in assigning scrambled gene segments from the micronucleus to a subset of genes from the macronucleus and allow us to study the general cellular and evolutionary mechanism of its global DNA rearrangements. (**2**) We request high quality draft sequence of the macronuclear genome of *Stylonychia lemnae*. *S. lemnae* is one of the most studied ciliates, and its evolutionary distance from *O. trifallax* is appropriate for efficient phylogenetic footprinting (see below for details). A comparative study between *S. lemnae* and *O. trifallax* genomes would lead to a better deciphering of both genomes in a synergystic manner, and lead to a better understanding of the genome unscrambling and DNA elimination mechanism involving small RNAs. (**3**) We suggest the sequencing of an additional five macronuclear genomes within the *O. trifallax* species complex, to generate data for phylogenetic shadowing. The proposed genomes include two additional species (*O. fallax* and *O. bifallax*) and three additional strains (*O. trifallax* JRB510, and provisionally JRA55 and JRB317, or a replacement to be determined after pilot sequencing). These sequences will be sufficiently similar to *O. trifallax* JRB310 that we can treat them as re-sequencing projects, using the high quality macronuclear genome sequence of *O. trifallax* 310 as the reference, to assemble inexpensive 454 generated data. Finally, (**4**) we request high quality draft sequence of the macronuclear genome of *Euplotes crassus*. Both *E. crassus* and *S. lemnae* have been important study systems for a hundred years or more, and currently have a community of molecular biologists using them as model systems for subjects such as chromatin remodeling,

telomere biology, translational frameshifting, and genetic code reassignment. Like *O. trifallax*, both E. *crassus* and *S. lemnae* have extensively fragmented macronuclear genomes of similar complexity to *O. trifallax*. Thus the ensemble of these spirotrich genomes creates an excellent model system for eukaryotic gene and regulatory sequence annotation, with most of the gene-finding already performed by the drastic DNA elimination events that form nanochromosomes. In addition, *Euplotes* is a deeply diverged spirotrich, whose evolutionary distance to the other spirotrichs (*Oxytricha* and *Stylonychia*) and oligohymenophora (*Paramecium* and *Tetrahymena*) makes it a good candidate species to bridge these two well-studied groups. For both *S. lemnae* and *E. crassus*, the *O. trifallax* macronuclear genome will most likely not serve as an effective aid in their assembly, so we suggest that these two genomes be sequenced at high quality draft coverage (5X), taking advantage of the full range of methods that we and our collaborators have explicitly developed for assembling and characterizing the *O. trifallax* genome. Also for these two species and for *O. fallax,* cDNA sequences (25,000 each) are requested. This would provide approximately 1X average coverage per coding gene, which is expected to cover only a fraction of coding genes, but would still provide a large dataset of confirmed gene structures for training computational gene-finding tools, thus making comprehensive macronuclear genome annotation more accurate.

Unlike *Paramecium* and *Tetrahymena*, the macronuclear genomes of *Oxytricha, Stylonychia* and *Euplotes* all have nano-chromosome structures formed by the most elaborate process of DNA elimination, genome rearrangement and chromosome breakage known in biology. The condensation of all cis-regulatory elements and coding regions into kilobase macronuclear nanochromosomes may be viewed as a sophisticated *biological* gene-finder. Nanochromosomes that lack conventional genes provide a good source for the discovery of novel classes of genes, including ncRNAs. A comparative study of these minimalist genome architectures that can pack over 30,000 genes in only 50 Mb will offer a unique power of resolution to zoom in on gene sequences and regulatory elements, with conclusions that are likely to extend to much larger eukaryotic systems.

Introduction

Ciliates are single-celled eukaryotes of fundamental biological interest and hold a special place in modern molecular biology. Ciliates were the birthplace of telomere biochemistry (e.g., Collins 1999). Cech (1990) discovered the first known self-splicing intron in *Tetrahymena*. Ciliates radically illustrate the C-value paradox (Gall 1981). Classically they were premier genetic eukaryotes in the hands of Tracy Sonneborn, his contemporaries, and his students (Sonneborn 1977, Nanney 1981).

Ciliates diverged from other microbial eukaryotes quite recently, as part of a monophyletic lineage (alveolates) with apicomplexans (e.g., *Plasmodium*) and dinoflagellates (Wright & Lynn 1997; Baldauf et al. 2000). Therefore, as a phylogenetic outgroup, ciliates provide a foil to studies of the crown eukaryotes: plants, animals, and fungi. Ciliates indeed are some of the best studied protists. Ciliate molecular genetics has been concentrated in two of the eleven ciliate classes, the oligohymenophorans, including *Tetrahymena* and *Paramecium*, and the spirotrichs, including some of the best studied ciliate species, in particular, the *Euplotes* lineage ("hypotrichs," Bernhard et al. 2001) plus several stichotrich species of the genera *Stylonychia* and *Oxytricha*. Each is quite diverged from the next and has its specific virtues; this diversity fuels comparative biology by a vigorous research community, focused on either the workings of ciliate germline/soma nuclear dimorphism, or on fundamental eukaryotic biochemical pathways and cellular processes.

The macronuclear genome. Each ciliate carries in its single cell two kinds of nuclei: 1) a typical diploid, meiotic germ-line nucleus (the micronucleus or **MIC**), and 2) a macronucleus (**MAC**), a highly differentiated organelle that provides all the transcripts for cell function. The MAC develops from a mitotic copy of the MIC, immediately after cell mating, haploid gametic nuclear exchange, and zygosis. Stichotrich MAC differentiation is the premier showcase of somatic genome alterations (review: Jahn & Klobutcher 2002), and has been a major focus of ciliate research since the discovery of these dramatic processes in *Stylonychia* by Ammermann (1974). MIC chromosomes polytenize, non-coding DNA sequences interrupting MAC-destined sequences are spliced out as DNA, the MDS gene segments are unscrambled, polytene chromatids are fragmented and telomeres are added *de novo* onto the new ends by telomerase, and finally the resulting MAC "chromosomes" are amplified to a level scaled to the size of the ciliate. The old MAC is apoptotically destroyed as the new MAC differentiates.

MIC-limited and MAC-destined sequences are interspersed in the MIC genome. In *Oxytricha* >90% of the sequences are MIC-limited, being destroyed in the developing MAC. The new MAC is entirely responsible for vegetative, clonal growth of the exconjugant (Herrick 1994). Introns are effectively both rare and extremely short (avg.=118 bp in stichotrichs) and subtelomeric non-coding sequences are short (Hoffman et al. 1995): i.e. there is on average 100bp separating both the start and stop codons from the telomere addition site (Cavalcanti et al. 2004a, 2004b). Thus, the MAC is nearly pure coding DNA.

Qualitatively, MAC differentiation is similar in most ciliates (Jahn & Klobutcher 2002, Yao et al. 2002), but the number of DNA splicing events, break sites, and chromosome kinds are orders of magnitude higher in spirotrichs than in oligohymenophorans. Oligohymenophoran MAC chromosomes contain hundreds of genes each, whereas spirotrich MAC chromosomes contain one or a few genes each, making the spirotrich MAC genome ideal for gene discovery. This spirotrich pattern seems to be shared with other less-studied classes of ciliates (Riley & Katz 2001), suggesting that massive fragmentation may be both ancestral and more representative of ciliate biology than the modest fragmentation in *Tetrahymena* and *Paramecium*.

MAC sequence complexity is surprisingly constant across the range of ciliates (most are ~50 Mb, although *Tetrahymena* has ~100 Mb), while ploidy levels scale with the size of the ciliate (Soldo et al. 1981, Ammermann & Muenz 1982). The mature *Oxytricha* MAC genome consists of genes deployed on a collection of ~25,000 different chromosomes amplified to an average ploidy of ~1000/MAC. These linear "chromosomes" are tiny, with an average length of ~2400 bp, ranging from ~0.25 to ~40 kb (Maercker et al. 1999); Figure 1 shows gel analysis of O. trifallax MAC DNA, and the distribution of kinds of chromosomes across the size range. Ciliates carry a large number of genes (≥30,000), compared to gene counts for *Drosophila melanogaster* (~14,000) and *Caenorhabditis elegans* (~18,500) (Rubin et al. 2000).

Internally eliminated sequences (IESs) and scrambled genes. During polytenization of the MIC genome, many sequences internal to MAC-destined sequences are precisely excised and eliminated (Internal Eliminated Sequences, or IESs; Klobutcher & Herrick 1997). Almost all spirotrich and *Paramecium* genes are interrupted by multiple IESs. IESs are of two types: some are cut-and-paste transposons (~4-5 kb long, with thousands of copies per haploid genome), while others are short (< 0.5 kb) AT rich non-coding sequences. Excision of IESs precisely reverts the germline insertional mutation, enabling expression of the gene in the MAC, and allowing the element to persist in the germline.

A related phenomenon is the exciting discovery of *scrambled genes* in the stichotrich MIC (reviewed in Prescott 2000). Again genes are interrupted by many IESs, but contiguous gene segments are no longer adjacent, nor necessarily on the same strand or even at the same locus in the MIC (Landweber et al. 2000), and the order of gene segments is permuted in both seemingly random patterns as well as clearly nonrandom patterns, like 1-3-5-7-2-4-6-8. IES removal permits linking of coding segments in the correct, translatable order and orientation. A well-studied example, the DNA polymerase alpha gene, is scrambled into 51 segments (Hoffman & Prescott 1997), dispersed on both strands of two unlinked MIC loci. While many cases are less scrambled, the Landweber lab has found at least one locus that is present in over one hundred scrambled MDSs (Maquilan et al., unpublished). As many as 20-30% of Oxytricha genes are estimated to be scrambled in the MIC genome. Traditionally, the study of scrambled genes has been laborious, and until recently most cases were discovered by chance. The modest amounts of MIC sequence coverage that we propose (2X) will provide enough data to reveal many hundreds of patterns and features of scrambled genes, permitting their rigorous study in an unbiased fashion.

How will additional spirotrich genomes inform our understanding of eukaryotic biology and human disease?

• Ciliates form a highly diverse clade of microbial eukaryotes that diverged just before the crown taxa of plants, animals, fungi, and stramenopile algae (Fig. 2), and their genes find homologs in crown eukaryotes, including metazoans, plants and fungi, equally well. The identification of protein products often depends on the availability of orthologs in other genomes: these are sequenced both directly and economically in large numbers from

spirotrich MAC DNA, because the DNA contains little more than coding regions plus short regulatory DNA (introns are few and small, and alternative RNA splicing has not been observed). By comparison, the estimated gene number of the compact vertebrate genome of *Takifugu rubripes*, originally selected based on its suitability for the detection of genes and regulatory elements (Aparicio et al. 1995), is roughly the same as that for *Oxytricha trifallax* (30,000), but the fugu genome is ~8X larger (400 Mb vs. 50Mb) (Aparicio et al. 2002). The macronuclear genomes of spirotrichs do not appear to contain "gene deserts" such as those found for fugu. The genomes of ciliates also have the potential to greatly facilitate computational and experimental detection of eukaryotic cis-regulatory elements, since only a tiny fraction of the *Oxytricha* macronuclear genome appears not to encode protein or RNA products, and this small fraction is the most likely fraction to contain regulatory elements. The "regulatory" genome portion of spirotrich macronuclear genomes, may be the smallest for known eukaryotes with a comparable gene complement. In our original pilot project (Doak et al. 2003; Cavalcanti et al. 2004a,b) 9.2% of the 553 reads with homologs in databases found a human homolog, but not one from yeast, fly, or worm (BLAST E-value ≤ e-15); extrapolating, ~2500 human genes have ciliate—but not yeast, worm, or fly—homologs. These genes are cases where ciliates may be the only non-vertebrate model system available.

• Because of their deep diversity, members of different classes of ciliates provide different—though complementary—information; thus, the MAC sequences of spirotrichs synergize strongly with oligohymenophoran gene sequences, providing broader datasets for comparison against their parasitic alveolate relatives and crown eukaryotes.

• Ciliate sequences will serve as a foil to the apicomplexan parasites (e.g., *Plasmodium*, *Toxoplasma*, and *Theileria*), their sister group within the alveolates. *Plasmodium,* with only ~5300 genes (Gardner et al. 2002), likely lost thousands of genes as it became host-dependent, perhaps in parallel gaining pathogenesis genes. Inferring events during the evolution of parasitism will aid in designing therapies. Furthermore, over 2500 protein-coding genes in *Plasmodium falciparum* are still un-annotated. This poses a very serious problem for understanding the parasite's genome, probably exacerbated by its high A-T content, making orthologous relationships more difficult to identify across distant taxa. By providing a broad set of close outgroups, the availability of many high quality and well-annotated gene-rich ciliate genomes should help annotate a portion of these genes in *Plasmodium*.

• As complex unicellular eukaryotes with elaborate cellular processes, ciliates are strong research organisms for cellular physiology. Their cellular organization is far more representative of metazoan cells than is that of yeast (Orias 2000), and their large gene sets reflect this.

• Discoveries essential to understanding human biology and health have repeatedly been made in ciliates: e.g., telomere structure and telomerase, histone modifications, and many aspects of cytoskeletal structure. Critical features of telomeres were first discovered in *Oxytricha*: in particular, the identity of all ~40 million MAC telomeres (Herrick & Wesley 1978; Klobutcher et al. 1981), the 3' G-strand overhang (Pluta et al. 1982; Klobutcher et al. 1981), and telomere-associated proteins and their interaction with the telomere sequence (Gottschling & Cech 1984; Gottschling & Zakian 1986; Classen et al. 2001). Multiple complete ciliate genomes will greatly facilitate the identification of complete enzymatic pathways and subcellular systems, allowing more comprehensive studies. Ciliates will surely continue to be key unicellular model organisms for the study of development, differentiation (especially gene/genome rearrangements), the cytoskeleton and cell motility, electrophysiology, and chromatin structure and function.

• Epigenetic phenomena have a long history in ciliate biology and a recent resurgence, with new discoveries of RNA's role in guiding chromatin remodeling and DNA rearrangement (e.g. Mochizuki and Gorovsky 2005; Juranek et al. 2005). An historical example is the maintenance/replication of cell-surface (cortical) structures and patterns (Sonneborn 1975). Even aberrant cortical patterns created by microsurgery can be maintained, leaving many questions open, such as pattern formation, organizing centers, propagation through cell division and even encystment. These issues are relevant to cytoskeletal function and the maintenance of cell differentiation in general. Recent studies have rekindled intense interest in Sonneborn's demonstrations that the genotype of the parental MAC can guide differentiation of the emerging MAC on a sequence-specific basis (Meyer & Garnier 2002). The epigenetic signals that pass from old to new MAC are now thought to be transferred via a novel twist

on an RNAi-type pathway. RNAi-based gene silencing also works on most ciliates (Bastin et al. 2001; Paschka et al. 2003) and is even relatively straight-forward for spirotrichous ciliates, because they will eat *E. coli*, particularly if supplemented with some of their regular food algae (Nowacki et al. in preparation).

• Many eukaryotic nuclear or mitochondrial genomes use alternative genetic codes, but the greatest code diversity of all is known in ciliates, where the use of either the standard code or one of four alternative codes is polyphyletic. For example, UAR encodes Gln in both oligohymenophorans and stichotrichs, while UGA encodes Cys in Euplotes but encodes Trp in two independent lineages. Possibly an evolutionary intermediate, the anaerobic ciliate *Nyctotherus* altogether avoids one of the three stop codons (Liang et al. 2005); tracing co-evolution of the code in ciliates and its associated translation components, such as release factors, currently offers the best model system for probing the selective and biochemical forces that drive code evolution (Lozupone et al. 2001).

• The biology of ciliate transposons is especially rich. The relationship of Mariner/Tc1 and Pogo-like transposases to retroviral integrases was first recognized because of new transposons discovered in ciliates (Doak et al. 1994). But in addition to transposase, ciliate transposons encode novel protein kinases and tyrosine recombinases (Doak et al. 1997; Doak et al. 2003). Also, the population dynamics of ciliate transposons is novel: genes evolve under a purifying selection for protein function (Klobutcher & Herrick 1997). This could be unique to ciliates, but more likely it is an extreme on a continuum of different transposon-host relationships. We expect that understanding the ciliate case will illuminate the coevolution of hosts—such as humans—and of their transposon parasites. Ciliate transposons are also the likely ancestors of IESs (see above), an evolutionary transition analogous to that proposed for Group II self-splicing introns as ancestors of splicesomal introns (Cavalier-Smith 1985).

Current status of the Oxytricha genome project.

Currently, we have a rather extensive analysis of a 5x coverage assembly of the macronuclear genome, and 200K additional reads are being generated. In addition, we have generated 1.3M 454 reads from a developmental series of cDNAs, isolated through conjugation and MAC development, plus ~1500 longer cDNA reads generated by the Canadian PEP project (Mike Gray and colleagues). This has allowed us to extend the preliminary analyses dating from earlier pilot projects (Cavalcanti et al. 2004a,b, Doak et al. 2003), and to monitor how our approaches to the fragmented MAC genome are progressing.

We have addressed three general questions: given the current coverage, 1) how completely has the macronuclear genome been sequenced; 2) how well is heterozygosity treated in the assembly; and 3) how severely do highly redundant sequences affect the sequencing effort? These issues are not entirely independent.

1) In the current assembly, 475,700 reads have been assembled into 50,014 contigs, which can be further grouped into 35,607 super-contigs based on homology, while we anticipate ~27,000 nanochromosome types. The consensus sequences of the current assembly sum to 75Mb, while the reported MAC sequence complexity is 50Mb, and there are relatively few singletons, indicating decent coverage. To assess the completeness of the genome, we used several different measurements. First, by the number of telomeres, we would expect at least 54,000 telomeres, and we now accounted for only 41,081 telomeres, so we are clearly lacking a large part of the genome, by this measure at least~20%. Second, we find that of 1069 EST clusters (>50bp), 951 can be mapped to WGS and singleton data, while 118 (~11%) cannot be mapped and are not likely to be contaminants. Third, Jung and Eddy find 93 tRNA genes, with 36 different anti-codons. Considering the wobble position, there are 12 (~18.8%) anti-codons missing. When singletons are included, the number of missing anti-codons decreases to 5 (~8%). In summary, these different measures suggest the current assembly is probably about 10~20% incomplete, but a new assembly incorporating 30% more reads is currently in progress.

2) The fact that the sequenced *Oxytricha* strain is a diploid wild isolate, heterozygous at many loci, presents both challenges to the genome project and offers additional information about the organism. At extremely high heterozygosity levels, the size of the genome to be sequenced effectively doubles, to 100Mb. Using a small but well-studied dataset of MAC sequences, the results showed that the divergence rate (including substitutions and indels) between highly similar MAC sequences (possible alleles or very recent paralogs) is on average ~2.2%, with

3.6% in the subtelomeric regions and 1.88% in the non-subtelomeric regions. As expected, we found that the assembler may segregate allelic or very recent paralogous sequence into different contigs when the divergence level is high enough (>3.5%), while the reads with lower divergence levels tend to be placed in common contigs. We conclude from this analysis that heterozygosity is inflating the effective size of the genome to be sequenced (the consensus sequences of the current assembly sum to 75Mb). This is therefore reducing the current effective coverage of each haploid genome., and more sequences will produce a better coverage of the haploid genome.

3) Hundreds of chromosomes are over-represented in the reads, presumably because they have a higher copy number in the macronucleus; however, these account for only 20% the total reads. There is no subgroup of highly repetitive sequences that result in a bi-modal distribution of coverage depth, and when we examined the identity of the top contigs, they shared no common features. There is no obvious way to exclude high copy number chromosomes, but nor is it a significant problem. Given that this is a diverse class of sequences, the only obvious way to eliminate them would be via Cot renaturation/fractionation. There are also many contigs that are underrepresented for reads – presumably the low copy nanochromosomes.

The need for additional comparative genome sequencing

In order to provide a broader spectrum of sequence conservation than that provided by *Drosophila melanogaster* and *D. pseudoobscura* alone (50 MyA divergence time), the genomes of *D. yakuba* and *D. simulans*, two species which have diverged comparatively recently from *D. melanogaster*, have been sequenced [see: Proposal for the Sequencing of *Drosophila yakuba* and *D. simulans* - Begun, Langley 2003]. Currently, *Paramecium* and *Tetrahymena,* two deeply diverging lineages within the Oligohymenophora, are the only two completely sequenced ciliate genomes (Aury et al. 2006, Eisen et al. 2006). However, preliminary ortholog group identification among *Paramecium*, *Tetrahymena* and *Oxytricha,* using only sequences from the latter's pilot project, suggests that only ~22% of the *Oxytricha trifallax* sequences have orthologs in either/both of the other two ciliates (see Figure 3). The DNA sequences are not alignable and the protein sequences align poorly between *Oxytricha* and the other two ciliates due to high divergence (see Figure 4). These results are not surprising given that both *Paramecium* and *Tetrahymena* diverged from *Oxytricha* >1 billion years ago (a distance similar to that between human and yeast). In addition, *Oxytricha's* MAC genome is more extensively fragmented and the chromosome copy numbers are more variable than that of either *Paramecium* or *Tetrahymena*, indicating that many basic cellular mechanisms may be highly diverged. Therefore, although the availability of *Paramecium* and *Tetrahymena* genome sequences enhances many aspects of ciliate research, to reach a better understanding of the nearly completed *O. trifallax* genome, comparative genomic studies among more closely related species are essential. In a similar manner to the *Drosophila* species, the addition of sequence data for relatively recently diverged *Stylonychia* as well as closely related *Oxytricha* species and *O. trifallax* isolates will enable identification of both rapidly and slowly evolving genes and other functional elements within these genomes. *Euplotes* will serve as a more distant relative and will assist in bridging the distance between *Tetrahymena* and *Paramecium* and the spirotrichs.

Proposed genome sequences

It is well-established that the rate of substitution of different genomic regions is not constant (for instance, Waterston et al. 2002); this is illustrated in the context of *Oxytricha* and related ciliates by Seegmiller et al. (1996), as well as the estimates of neutral rates we present in Table 2. Variability in genomic substitution rates leads to trade-offs between the ability to align different genomic portions and the ability to detect sequence motifs (Eddy 2005; Stone, et al. 2005). These trade-offs can be overcome by careful selection of genomes of organisms at appropriate evolutionary distances from each other. For the purpose of exploring some of these trade-offs in the context of the genomes we propose to sequence, we will refer to estimates shown in Table 2.

1. Proposed additional *O. trifallax* micronuclear sequence

The originally approved *Oxytricha* whitepaper included sequencing a portion of the MIC genome equivalent to the MAC genome, that is, 50Mb. The total MIC sequence complexity is ~1Gb, so this original modest proposal requested only 1X coverage of 5% of the genome. This was with the goal of sequencing a small portion of the

MIC loci that give rise to MAC chromosomes, to generate an unbiased view of the extent of gene scrambling in the MIC. We are currently generating libraries for this project and have performed survey sequencing of some MIC clones, but, five years after the original *Oxytricha* whitepaper proposal, we would like to request the addition of slightly deeper coverage of the *Oxytricha* MIC, to permit reasonable accuracy in assigning MDSs (gene segments) from the MIC to a subset of genes from the MAC. Our simulations (Yi Zhou et al.) suggest that 2X coverage of the MIC genome would permit recovery of ~70% of the MDSs for at least one allelic type (whereas we can only expect to recover <5% in the currently proposed 50Mb survey coverage). In addition, depending on the sequencing strategy (fosmid or shot-gun), at 2X coverage, roughly 25%-50% of the genes can be almost (>85% of their length) completely covered in the MIC sequence. (Simulations were based on the expectations that the MDS portion is 5% of the MIC, and that roughly 20% of the intragenic regions are IES.)

We suggest that the 2X MIC sequence be produced by a standard combination of shattered shotgun libraires, and fosmid or BAC clones. Up to now, we have concentrated on generating BAC or fosmid clones of MIC DNA, which has been difficult, given problems in generating large amounts of megabase sized MIC DNA. We remark that only 1% of the cellular DNA is MIC, and the rest is MAC, so that even a 100x purification strategy leaves 50% MAC contamination, and handling leads to additional breakage of MIC chromosomes. We are now working closely with Lucigen to produce "small insert" BAC libraries, and are optimistic that this will yield usable libraries very soon. In addition, other approaches in the Landweber lab have yielded several MIC clones in the 10-20 Kb range, currently in the sequencing and assembly pipeline. Shatter libraries of the MIC will be easier to generate: the DNA doesn't need to be intact, and a small level of MAC contamination will just contribute to the MAC assembly effort by providing resequencing (see the companion *Tetrahymena* comparative genomics whitepaper).

2. Proposed macronuclear sequence of *Stylonychia lemnae*, both as an important experimental model and for phylogenetic footprinting

Like *Oxytricha*, *Stylonychia* is a stichotrich, and its biology is thus comparable to that of *Oxytricha*. Because it is physically larger, it was used extensively in early cytogenetic studies of macronuclear development. More recently, it has been used to develop RNAi and transformation methods, and as a comparative model for gene scrambling – it shares many scrambled genes with *Oxytricha*, but their scrambled germline architectures can strikingly differ (Möllenbeck et al. 2006). We collected 80 *S. lemnae* protein sequences and 115 cDNA sequences that are non-redundant from the NCBI database. Approximately 64% of the *S. lemnae* cDNA sequences have blastn hits with E-value<1e-3 in the current *O. trifallax* macronuclear genomic DNA assembly. The alignable regions usually extend to most of the sequence lengths and the average identity level peaks at 70% (as seen in Figure 5, A, where identity level=[matched bp/cDNA length]). Almost all of the *S. lemnae* protein sequences (~95%) find blast hits (E-value<1e-3) in *O. trifallax*. The proteins in these two species are highly similar (identity level peaks at ~90%, as seen in Figure 5, B). These preliminary analyses, together with the estimates of neutral evolutionary distance listed in Table 2, suggest that the divergence between *Oxytricha* and *Stylonychia* is appropriate for phylogenetic footprinting (Eddy 2005), which can efficiently identify longer functional segments (~50bp) including protein-coding genes and non-coding RNAs. Furthermore, the nanochromosome architecture in *Stylonychia* and *Oxytricha* compresses all cis-regulatory elements into typically less than **150bp** subtelomeric regions (~100 bp average) flanking each gene sequence, effectively shrinking the search space for phylogenetic footprinting and enabling the identification of even smaller functional motifs, such as those that regulate RNA transcription or DNA copy-number in the macronucleus (see Figure 6).

3. Proposed Macronuclear sequence of *O. fallax* and multiple *O. trifallax* strains, for phylogenetic shadowing.

Although we expect the comparative study between the *S. lemnae* and *O. trifallax* genomes to help us identify most of the functional units in both species, additional effort is needed to produce a high quality annotation of the *Oxytricha* genome, to resolve alleles in the current genome assembly, and to identify other unique biological features that are specific to *Oxytricha*.

Following Eddy (2005), we propose to re-sequence five additional genomes in the *Oxytricha* species complex to provide enough data for phylogenetic shadowing. All of the proposed genomes in this set are roughly 5%

divergent at the DNA sequence level from *Oxytricha trifallax* and should be easily aligned to the *O. trifallax* assembly, therefore; they are appropriate for "resequencing" projects, relying primarily on 454 pyrosquencing or other current technology. Compared to conventional sequencing, a 454 approach currently reduces the cost per base ~9-fold, making it a cost-effective method to gather data for phylogenetic-shadowing. To our benefit, the compression of cis-regulatory elements onto nanochromosomes may dramatically reduce false positive levels relative to those for genomes with larger conventional chromosomes. We hope that this will lead to the requirement of fewer genomes for phylogenetic shadowing. In addition, the partitioning of the macronuclear genome into nanochromosomes may facilitate data aggregation strategies, resulting in a higher effective sequence element length. Therefore, these five genomes should not only provide statistical power to help us analyze large conserved sequences (>50 nt, like exons) similar to the comparison to *Stylonychia*, but we also hope they will provide a glimpse at shorter conserved sites (~10bp, like transcription-factor binding sites) because of the reduced space for regulatory sequences.

The *trifallax* species complex was originally identified in a search for *O. fallax*-like representatives (Bob Hammersmith, Ball State University). The large collection of cell lines that we call *Oxytricha trifallax* (most also fall in the recently renamed group *Sterkiella histriomuscorum*; Foissner and Berger, 1999) were isolated from over 500 individual strains from different limonitic sites in Indiana, during fall-winter 1985-6. ~60 clones obtained were morphologically indistinguishable from *O. fallax*. By pairwise crosses these were divided into a small number of mating complexes. Following Sonneborn's example of *Paramecium aurelia* syngens, these mating complexes were provisionally given species names *O. bifallax*, *trifallax*, etc. These isolates were further characterized at the molecular level via dot blots against *O. fallax* (Hammersmith and Herrick, unpublished). More recently, a subset of these were characterized by full-length small- and large-subunit rDNA sequencing (Doak et al., in preparation). Phylogenetic analysis of both datasets indicates that *O. fallax* and *trifallax* are closely related species, and that there is a fair degree of variation among *O. trifallax* isolates (see Figure 7).

Allelic divergence within *O. trifallax* is at the appropriate distance for phylogenetic shadowing. Seegmiller et al. (1996) specifically looked at the divergence between alleles within and between *O. fallax* and *trifallax* in coding, non-coding, and MIC limited sequences, for a specific locus. They estimated intraspecies (i.e. interallelic) divergence, D_S, to be <0.1 at synomomous sites and non-coding divergence, D, to be <0.2 at non-coding sites, while interspecies divergences were ~0.2 and ~0.4 respectively. These intraspecies divergences are an order of magnitude greater than those estimated for *Ciona intestinalis*, and two orders of magnitude greater than estimates for *Homo sapiens* (Dehal et al. 2002; Boffelli et al. 2004). The divergence between alleles within trifallax species therefore suggests that they are suitable for phylogenetic shadowing.

While we have *trifallax* isolates from more than a dozen geographical locations, JRB510 (isolated from the same location as JRB310, the strain being used for the current genome sequence) does not share alleles with JRB310 at the few loci examined, and RNA from JRB310 x JRB510 crosses have been used for all *Oxytricha* EST and cDNA sequences (as well as all matings used in the literature), such that the addition of the JRB510 genomic sequence will provide cognate genomic sequence for all mRNA/EST sequences available. Therefore we suggest that JRB510 is certainly our first choice of *O. trifallax* strain to use for resequencing. In addition, we would like to sequence at least 2-5 more *O. trifallax* strains, taken from widely separate locations, with precise strain selections determined by pilot sequencing to confirm allelic distances.

We also include *O. fallax*, a sister species to *O. trafallax*, in the dataset for shadowing. *O. fallax* is by far the closest species to *O. trifallax*, and historically *O. trifallax* was even isolated in an effort to reisolate *O. fallax*. All other species named *Oxytricha* are very divergent–clearly the *Oxytricha* genus is polyphyletic, which is generally recognized. The divergence between *O. fallax* and *trifallax* (3-15%, see Seegmiller et al. 1996 and Table 2) is optimal for phylogenetic shadowing. (The divergence between *O. fallax* and *trifallax* was estimated to be about 0.2 in Seegmiller et al., and about 0.05 in our estimate. This may simply reflect that their locus is more divergent than those we present in Table 2.) We request pilot sequencing to confirm whether *O. bifallax* is at the appropriate distance for phylogenetic shadowing, and we would be open to replacing this species with another taxon, even *O. nova*, if a greater neutral distance is sought, since *O. nova* is a current laboratory model for telomere biology.

4. Proposed macronuclear sequence of *Euplotes*.

Since researchers collaborating at Princeton and at the GSC at Wash U have developed successful approaches for sequencing ciliate macronuclear genomes with nanochromosomes typically less than 10 kb (usually just one gene on an approximately 2 kb chromosome), the opportunity is ripe to extend all of the molecular approaches and bioinformatic methods that we have developed for *Oxytricha* (including an *Oxytricha*-specific gene finder from Sean Eddy's lab) to a small group of other ciliates with similar genome architecture but different biological features that make them attractive to study. Together, Wash U, Princeton, and the Eddy lab have developed a strong arsenal of molecular and bioinformatic tools for sequencing and interpreting spirotrichous ciliate genomes.

In addition to the *Oxytricha* and *Stylonychia* species, *Euplotes* has also played a pivotal role in spirotrich research. Of the various *Euplotes* species studied, *Euplotes crassus* has been the most frequently used subject. *Euplotes* is the organism where telomerase was originally discovered (Lingner et al. 1997) and continues to serve as a very active model organism for telomerase studies (e.g. Fouche et al. 2006). Furthermore, in conjunction with other ciliates, *Euplotes* has led to characterization of the reverse transcriptase activity of this enyzme (Lingner and Cech 1996; Linger et al 1997), and the proposal of universal telomerase activity in all eukaryotes (Bryan et al. 1998). In addition to the use of alternative codes within the euplotids (see below), as many as 10% of its genes contain programmed frameshifts. The excision of both short and transposon IESs is well studied, but like *Paramecium* and *Tetrahymena* there is no indication of micronuclear genome scrambling to date.

While *Tetrahymena*, *Paramecium*, *Oxytricha*, and *Stylonychia* share the same UAR=Gln stop codon assignment in their genetic code, *Euplotes* has evolved one or more orthogonally distinct variants of the code, reassigning the stop codon UGA to an amino acid instead (and there are hints that *E. focardii* may also reassign UAG; Miceli et al. 1994). This in itself makes a *Euplotes* genome attractive to sequence, since no organism with this genetic code has been sequenced yet, and there is a tremendous interest in understanding the rewiring mechanisms that lead to genetic code change and the resulting impact it has upon the entire genomic landscape. Perhaps reflecting its possible status as an evolutionary intermediate state in genetic code change, *Euplotes crassus* displays frequent use of alternate translational decoding, including translational frameshifting. Its abundant reassignment of UGA from stop to glutamine represents only one of numerous independent lineages containing stop codon reassignments in ciliates; therefore the comparative genomics of *Euplotes* and its relatives (that use UGA as stop but use UAR as glutamine) provide an excellent model system to understand the evolutionary transitions between genetic codes. Secondly, ~10% of *Euplotes* genes require a +1 translational frameshift to produce the correct protein product (Klobutcher 2005). While programmed translational frameshifting is observed in a wide range of organisms, its abundance in *Euplotes* is unprecedented. Availability of the genome will help determine whether a specialized mechanism evolved to facilitate frameshifting in *Euplotes* or whether other aspects of *Euplotes* genetics have somehow fostered a tolerance for—or relaxed negative selection against—genes that require frameshifts for their expression.

A high quality draft of a *Euplotes* genome will uniquely allow inference of the global patterns of codon reassignment and +1 translational frameshifting in this genome, as well as whether particular sequence contexts are necessary for either or both of these processes. More generally, knowledge of the gene products encoded in the genome will aid in developing testable models, and, as in other model systems, availability of the sequences of relevant genes will permit experiments to test models. Developing *Euplotes* as a model system to study programmed ribosomal frameshifting, which is also found in other Eukaryotes (Namy et al. 2004; Baranov et al. 2002), and especially their retroviruses (Brierley 1995; Dos Ramos et al. 2004), will be an important impact of the genome sequence. Euplotes has also been developed as a genetic system, including artificial MAC chromosomes (Bender et al 1999; Erbeznik et al. 1999) and RNAi (Möllenbeck et al. 2003).

In the context of comparative genomics, the distance between *E. crassus* and *O. trifallax* is similar to that between *Paramecium* and *Tetrahymena*. Hilary Morrison and Mitch Sogin at the Marine Biological Laboratory have sequenced 6000 ESTs (~1800 clusters) from vegetative mRNA provided by Larry Klobutcher. Our preliminary analysis of these *E. crassus* EST data shows that in most coding regions *E. crassus* and *O. trifallax* are unalignable at the DNA sequence level (see Figure 8, A.). Only 16% of the *E. crassus* EST sequences have blastn hits with E-value<1e-3 in the *O. trifallax* macronuclear genomic DNA assembly. The alignable regions are usually small and

highly divergent (as seen in Figure 8, A. where identity level=[matched bp/EST length]). However, more *E. crassus* sequences (~60%) find blast hits (E-value<1e-3) in the *O. trifallax* genome at the protein sequence level, with identity level around 45% (as seen in Figure 8, B.). This identity level may be an overestimate due to the lack of knowledge of full protein lengths in either species. Although the average divergence level of *Euplotes* is beyond the regular range for phylogenetic footprinting, its genome is possibly even more fragmented than either *Oxytricha* or *Stylonychia*, resulting in smaller average nanochromosome size, and more extensive delimitation of functional units on the nanochromosomes. These features makes it possible to use comparisons among *Euplotes*, *Stylonychia* and *Oxytricha* to detect highly conserved functional elements in spirotrichs (see Figure 6 (rDNA footprinting)).

In addition, since *Euplotes* is a deeply rooted spirotrich, the sequencing of a *Euplotes* species would provide a broader species coverage in ciliates to help "bridge" the two best-studied ciliate groups, Oligohymenophorea (*Paramecium* and *Tetrahymena*) and Spirotrichea (*Euplotes*, *Stylonychia* and *Oxytricha*), the only two classes with well developed genetic and molecular tools. The inclusion of *Euplotes* in ortholog group analysis will help reach a much higher resolution within the ciliate/alveolate clade, revealing important events that might have previously been undetected in the ancestral lineages among these alveolates, and this will assist identification of orthologs in *Plasmodium*.

Figure and Table Legends

Fig. 1. Estimated distributions of MAC chromosome sizes for *O. trifallax*, *S. lemnae*, *E. crassus*, and *E. aediculatus*. Chromosome sizes were estimated from densitometry of EtBr post-stained whole cell DNA gel electrophoresis lanes. Background fluorescence was subtracted before normalizing raw image pixel intensities by the total sample DNA concentration, and chromosome length. The left graph indicates intensity following normalization for total sample DNA concentration; the right graph is fully normalized. Estimates of chromosome length were calculated for the region corresponding to 160 to 750 pixels from the wells, to exclude the MIC fraction and possible degradation products. 95% confidence interval levels (indicated by ±) were estimated via a boostrapping approach. The median chromosome length for O. trifallax is in good agreement with estimates derived from full-length chromosomes from the current *O. trifallax* coverage. The range for *E. aediculatus* may be slightly underestimated, because of a different method of purification.

Fig. 2. Concatenated protein tree for selected ciliates and reference species. Modified from Baldauf et al. (2000).

Fig. 3. Preliminary ortholog group identification among *Paramecium*, *Tetrahymena* and *Oxytricha*.

Fig. 4. Protein alignment and maximum likelihood tree for vacuolar ATP synthase orthologs.

Fig. 5. Distance comparisons between *S. lemnae* and *O. trifallax*. We collected 80 *S. lemnae* protein sequences and 115 cDNA sequences that are non-redundant from the NCBI database. Approximately 64% of the *S. lemnae* cDNA sequences have blastn hits with E-value<1e-3 in the *O. trifallax* macronuclear genome assembly. The alignable regions usually extend to most of the sequence lengths and the average identity level peaks at 70% (as seen in A, where identity level=[matched bp/cDNA length]). Almost all of the *S. lemnae* protein sequences (~95%) can find blast hits (E-value<1e-3) in the *O. trifallax* genome. The proteins in these two species are highly similar (identity level peaks at ~90%, as seen in B). The average divergence level between *S. lemnae* and *O. trifallax* is somewhat comparable to that between mouse and human or *C. elegans*/*C. briggsae*, making *S. lemnae* a good choice for phylogenetic footprinting.

Fig. 6. Identification of a 10-bp motif in spirotrich rDNA 5'-flanking sequence using footprinting

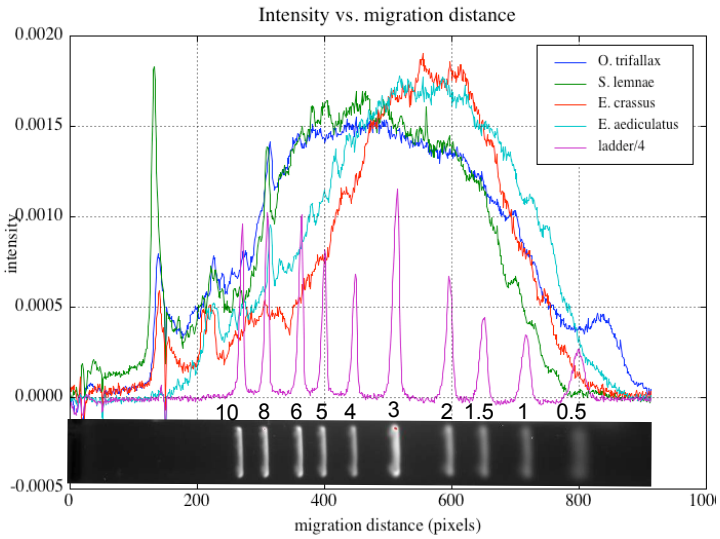
Fig. 7. Small-subunit rRNA PAUP Neighbor-joining tree for *O. trifallax* isolates. Bootstrap % shown.

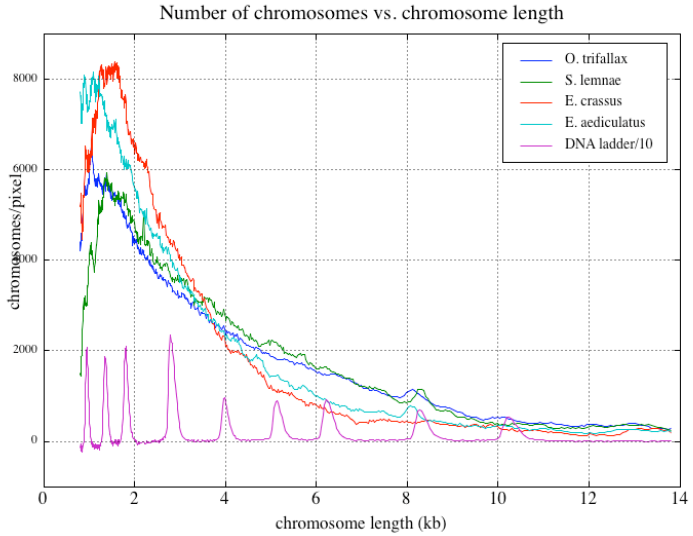
Fig. 8. Distance comparisons between *E. crassus* and *O. trifallax*. Our preliminary analysis of 1820 *E. crassus* EST sequence clusters shows that for most coding regions the two species are mostly unalignable at the DNA sequence level. Only 16% of the *E. crassus* EST sequences have blastn hits with E-value<1e-3 in the *O. trifallax* macronuclear genome assembly. The alignable regions are usually small and highly divergent (as seen in A, where identity level=[matched bp/EST length]). However, more *E. crassus* sequences (~60%) have blast hits (E-value<1e-3) in the *O. trifallax* genome at the protein sequence level, with identity levels near 45% (as seen in B). This identity level may be overestimated due to the lack of knowledge of full protein lengths in either species.

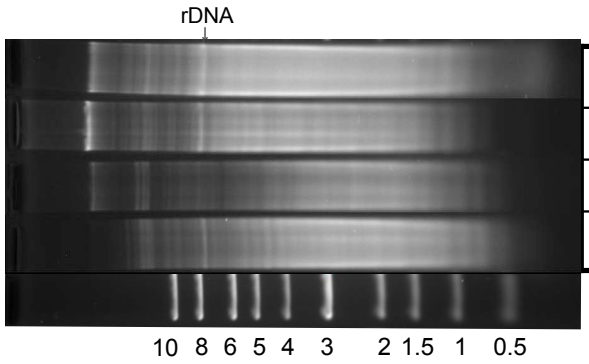
Table 2. Neutral divergence data (mean ± standard deviation) for pairwise alignments of orthologous coding and noncoding regions. For protein coding genes, the divergences were calculated from comparable 4-wobble codon positions of HSP70; eukaryotic release factor 1 (eRF1); alpha telomere binding protein (alpha-TBP); phosphoglycerate kinase (PGK); and ribosomal protein S21. The noncoding region included is the 5'-flanking sequence upstream of the ribosomal DNA transcription starting site (rDNA 5'-TSS). All divergence data were calculated relative to *O. trifallax* 310, and are corrected using the Kimura two-parameter model. Any of the alignments with a divergence level >75% should be treated with caution due to complete saturation.

[footnote 1: There are at least two cases where the neutral divergence rates for synonymous sites have been found to be substantially higher than those obtained for noncoding sites (Subramanian and Kumar 2003; Neafsery, Hartl, and Berriman 2005). If the situation is similar for the organisms we propose to sequence, then the estimates based on calculations from the coding sequence regions used may be too high.]

Figure 1







	Median length
O. trifallax	1.8 kb ± 0.16
S. lemnae	2.0 kb ± 0.17
E. crassus	1.7 kb ± 0.12
E. aediculatus	1.6 kb ± 0.13

Figure 2

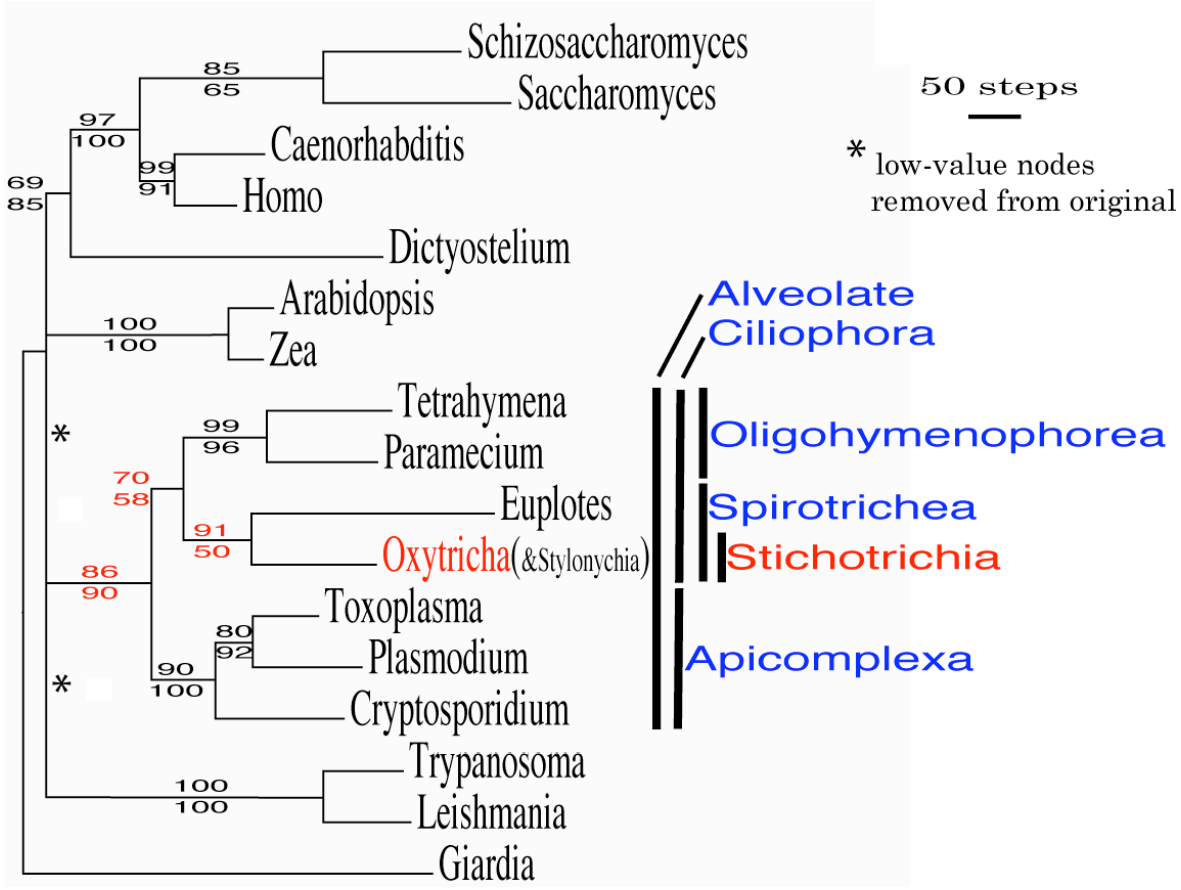

Schizosaccharomyces
Saccharomyces
Caenorhabditis
Homo
Dictyostelium
Arabidopsis
Zea
Tetrahymena
Paramecium
Euplotes
Oxytricha(&Stylonychia)
Toxoplasma
Plasmodium
Cryptosporidium
Trypanosoma
Leishmania
Giardia
85
65
97
100
99
91
69
85
100
100
99
96
70
58
91
50
86
90
80
92
90
100
100
100
*
*
50 steps
* low-value nodes
removed from original
Alveolate
Ciliophora
Oligohymenophorea
Spirotrichea
Stichotrichia
Apicomplexa

Figure 3

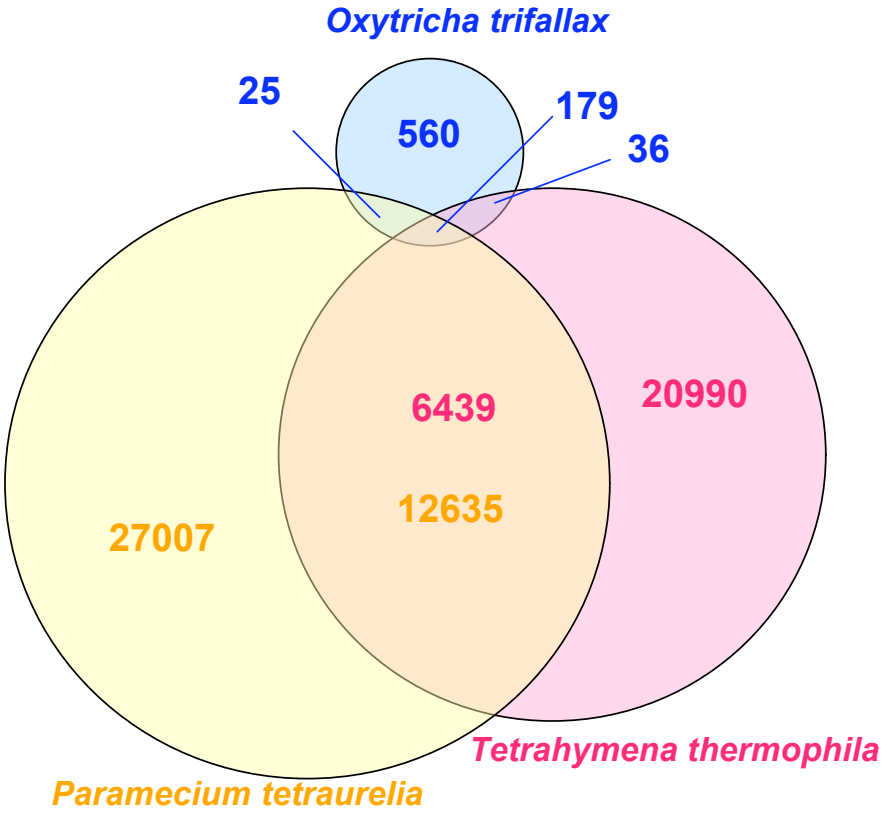

Oxytricha trifallax
25
560
179
36
6439
20990
12635
27007
Tetrahymena thermophila
Paramecium tetraurelia

Figure 4

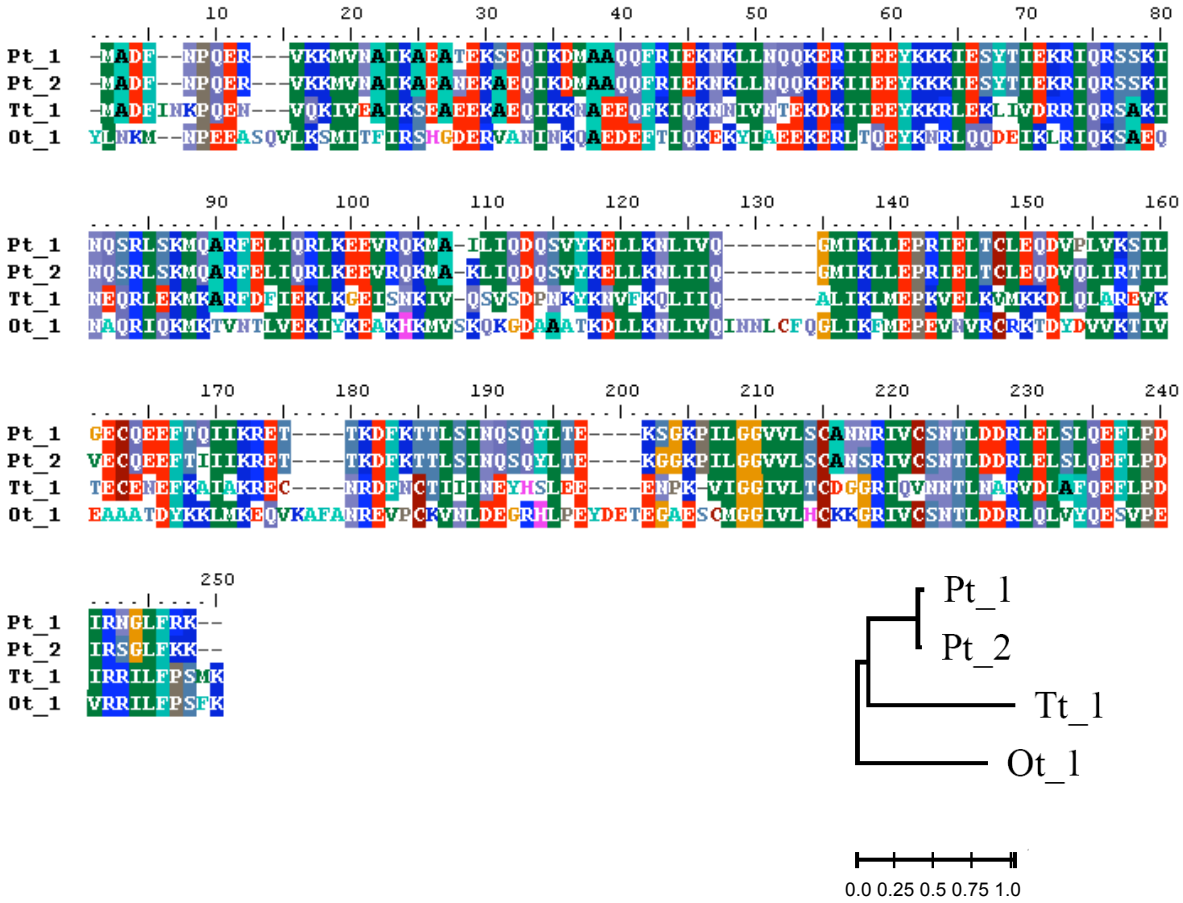


Figure 5

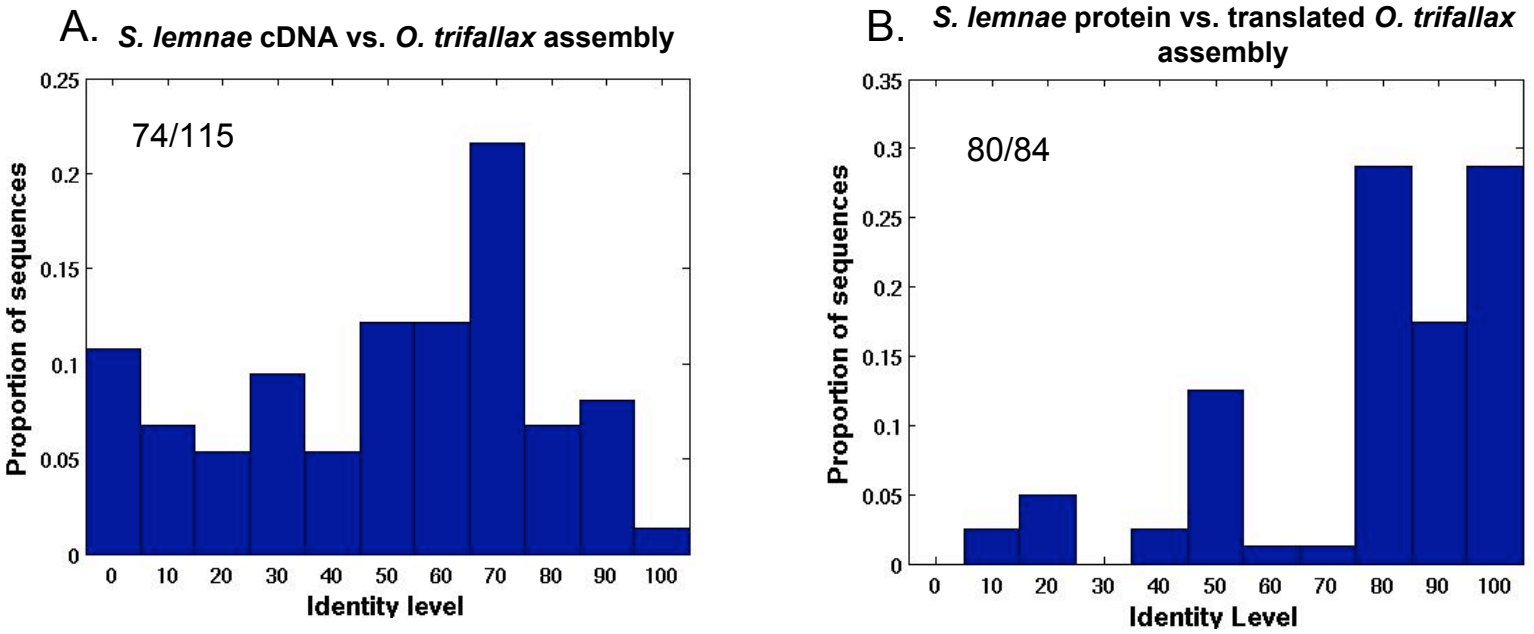
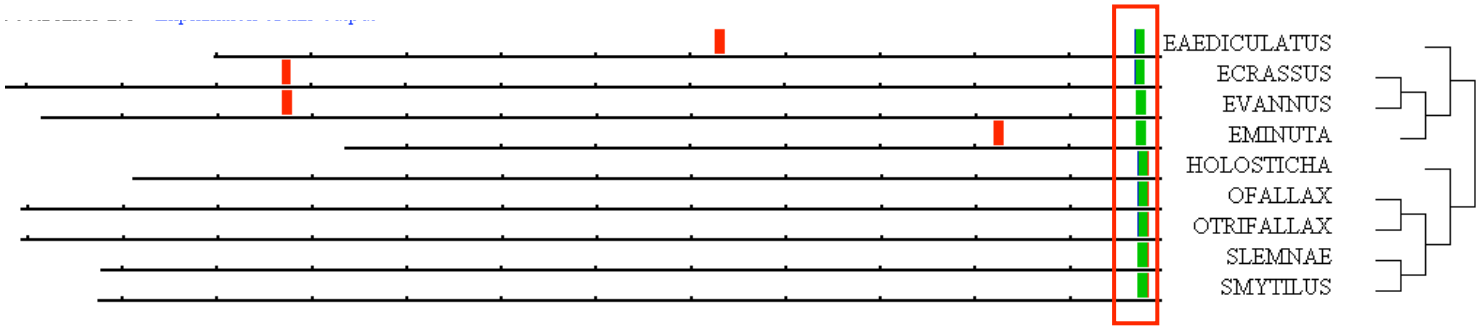

A. S. lemnae cDNA vs. O. trifallax assembly
74/115
Proportion of sequences
0.25
0.2
0.15
0.1
0.05
0
0 10 20 30 40 50 60 70 80 90 100
Identity level
B. S. lemnae protein vs. translated O. trifallax assembly
80/84
Proportion of sequences
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 10 20 30 40 50 60 70 80 90 100
Identity Level

Figure 6



Species	Motif position	Motif sequence
E. aediculatus	-30	AAAAGTTACG
E. crassus	-30	AAAAGTTACA
E. vannus	-30	AAAAGTTACA
E. minuta	-29	AAAAGTTACA
Holosticha	-28	AACTCTTACA
O. fallax	-28	AACTCTTACA
O. trifallax	-28	AACTCTTACA
S. lemnae	-28	AAATCTTACA
S. mytilius	-28	AAATCTTACA

Figure 7

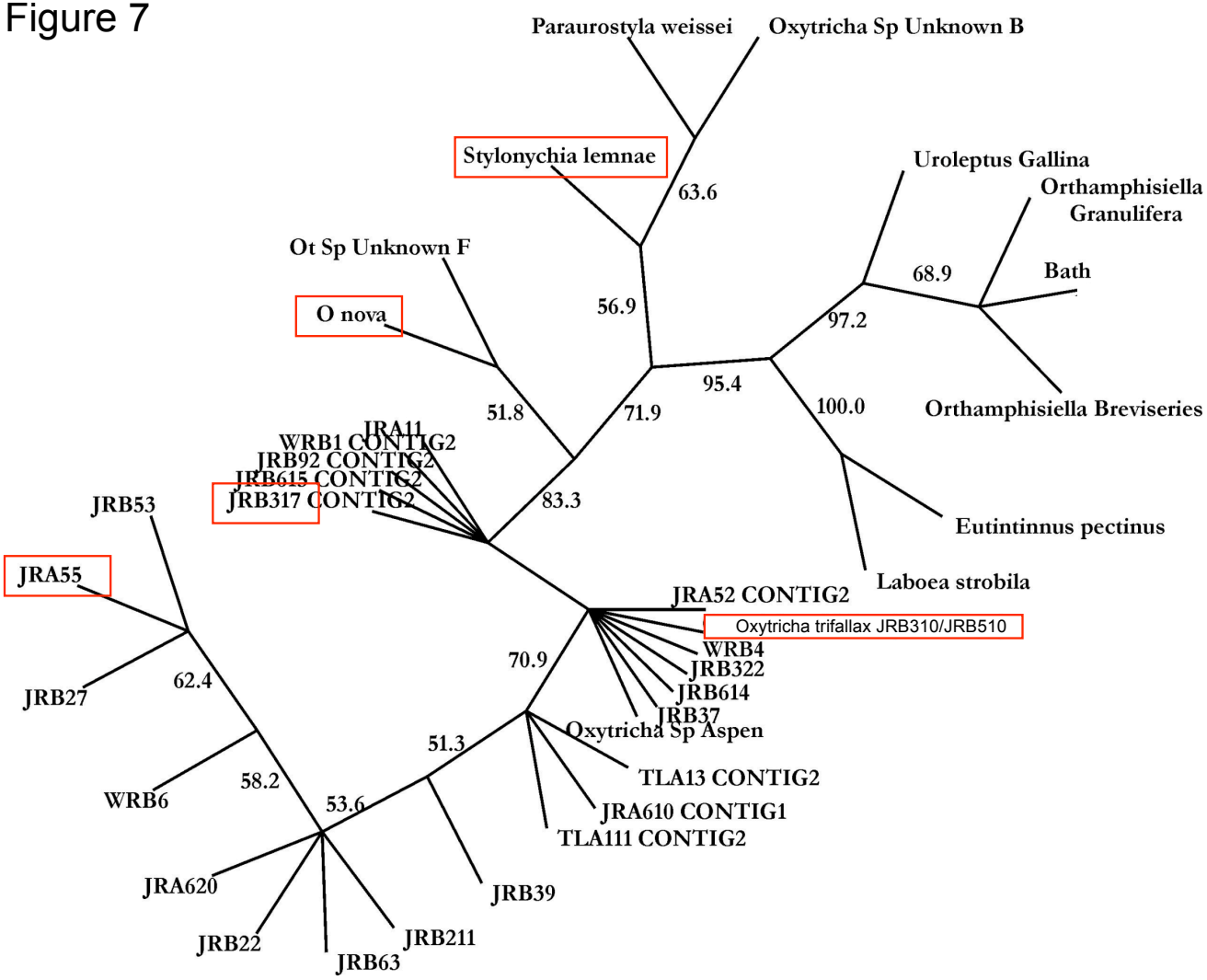
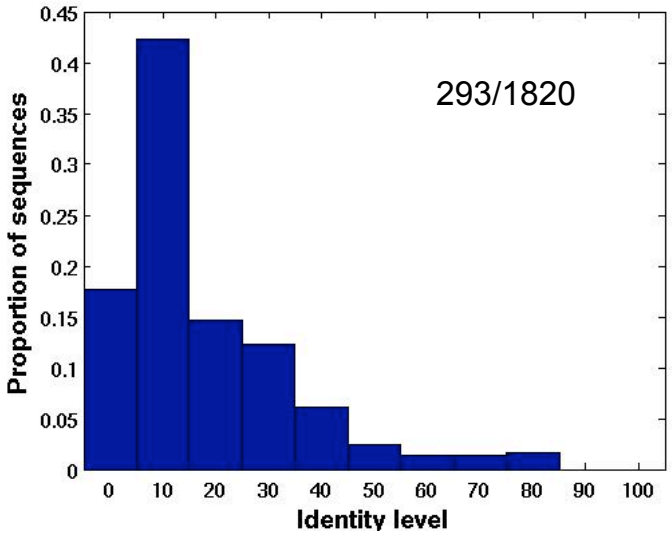

Paraurostyla weissei
Oxytricha Sp Unknown B
Stylonychia lemnae
63.6
Uroleptus Gallina
Orthamphisiella Granulifera
Ot Sp Unknown F
68.9
Bath
O nova
56.9
97.2
95.4
51.8
71.9
100.0
Orthamphisiella Breviseries
JRA11
WRB1 CONTIG2
JRB92 CONTIG2
JRB615 CONTIG2
JRB317 CONTIG2
83.3
JRB53
Eutintinnus pectinus
JRA55
Laboea strobila
JRA52 CONTIG2
Oxytricha trifallax JRB310/JRB510
WRB4
70.9
JRB322
JRB614
JRB37
62.4
JRB27
Oxytricha Sp Aspen
51.3
TLA13 CONTIG2
58.2
WRB6
53.6
JRA610 CONTIG1
TLA111 CONTIG2
JRA620
JRB39
JRB22
JRB211
JRB63

Figure 8

A. ***E. crassus* EST vs. *O. trifallax* assembly**



B. **Translated *E. crassus* EST vs. translated *O. trifallax* assembly**

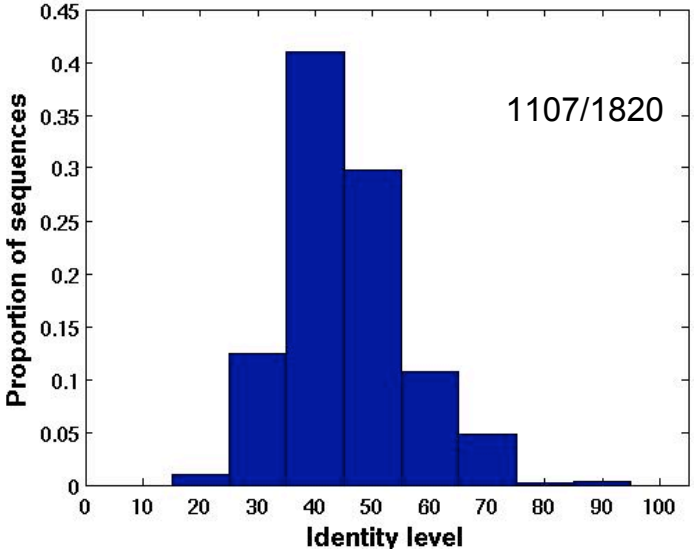


Table 2

	HSP70	eRF 1	alpha-TBP	PGK	S21	rDNA (5'-TSS)
Oxytricha fallax	-	-	-	-	0.036 ± 0.026	0.052 ± 0.006
Stylonychia lemnae	0.40 ± 0.05	0.61 ± 0.10	0.88 ± 0.15	-	0.16 ± 0.06	0.75 ± 0.04
Stylonychia mytilus	-	0.67 ± 0.23	0.66 ± 0.10	-	0.16 ± 0.06	0.75 ± 0.04
Oxytricha nova	0.71 ± 0.08	-	0.88 ± 0.15	0.91 ± 0.17	0.056 ± 0.032	-
Paraurostyla weissei	-	0.86 ± 0.16	0.77 ± 0.12	-	-	-
Paramecium tetraurelia	0.66 ± 0.09	-	-	0.83 ± 0.18	-	-
Tetrahymena thermophila	1.22 ± 0.26	-	-	1.12 ± 0.28	-	-
Euplotes crassus	0.75 ± 0.10	-	1.27 ± 0.51	1.22 ± 0.30	-	0.82 ± 0.05
Euplotes octocarinatus	-	0.93 ± 0.22	-	-	-	-

References

Ammermann D, Muenz A. 1982. DNA and protein content of different hypotrich ciliates. Eur J Cell Biol 27:22-24

Ammermann D, Steinbrück G, von Berger L, Hennig W. 1974. The development of the macronucleus in the ciliated protozoan *Stylonychia mytilus*. Chromosoma 45:401-429.

Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, Krumlauf R, Brenner S. 1995. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. Proc Natl Acad Sci U S A. 92:1684-8

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MD, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJ, Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science. 297:1301-10

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N, Arnaiz O, Billaut A, Beisson J, Blanc I, Bouhouche K, Camara F, Duharcourt S, Guigo R, Gogendeau D, Katinka M, Keller AM, Kissmehl R, Klotz C, Koll F, Le Mouel A, Lepere G, Malinsky S, Nowacki M, Nowak JK, Plattner H, Poulain J, Ruiz F, Serrano V, Zagulski M, Dessen P, Betermier M, Weissenbach J, Scarpelli C, Schachter V, Sperling L, Meyer E, Cohen J, Wincker P. 2006 Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature. 444:171-8.

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972-977.

Baranov PV, Gesteland RF, Atkins JF. 2002 Recoding: translational bifurcations in gene expression. Gene. 286:187-201.

Boffelli D, Weer CV, Weng L, Lewis KD, Shoukry MI, Pachter L, Keys DN, Rubin EM. 2004. Intraspecies sequence comparisons for annotating genomes. Genome Res. 14:2406-11

Bastin P, Galvani A, Sperling L. 2001. Genetic interference in protozoa. Res Microbiol 152:123-129.

Bender J, Kampfer M, Klein A. 1999. Faithful expression of a heterologous gene carried on an artificial macronuclear chromosome in *Euplotes crassus*. Nucleic Acids Res. 27:3168-3172.

Bernhard D, Stechmann A, Foissner W, Ammermann D, Hehn M, Schlegel M. 2001. Phylogenetic relationships within the class Spirotrichea (Ciliophora) inferred from small subunit rRNA gene sequences. Mol. Phylogenetics Evol. 21:86-92.

Brierley I. 1995 Ribosomal frameshifting viral RNAs. J Gen Virol. 76:1885-92.

Bryan TM, Sperger JM, Chapman KB, Cech TR. Telomerase reverse transcriptase genes identified in *Tetrahymena thermophila* and *Oxytricha trifallax*. 1998 Proc Natl Acad Sci U S A. 95:8479-84

Cavalcanti AR, Dunn DM, Weiss R, Herrick G, Landweber LF, Doak TG. 2004a. Sequence features of *Oxytricha trifallax* (class Spirotrichea) macronuclear telomeric and subtelomeric sequences. Protist. 155:311-22.

Cavalcanti AR, Stover NA, Orecchia L, Doak TG, Landweber LF. Coding properties of *Oxytricha trifallax* (*Sterkiella histriomuscorum*) macronuclear chromosomes: analysis of a pilot genome project. Chromosoma. 2004b. 113:69-76.

Cavalier-Smith T. 1985. Selfish DNA and the origin of introns. Nature. 315:283-284.

Cech TR. 1990. Self-splicing of group I introns. Annu Rev Biochem. 59:543-568.

Classen S, Ruggles JA, Schultz SC. 2001. Crystal structure of the N-terminal domain of *Oxytricha nova* telomere end-binding protein alpha subunit both uncomplexed and complexed with telomeric ssDNA. J Mol Biol. 314:1113-1125.

Collins K. 1999. Ciliate telomerase biochemistry. Annu Rev Biochem 68:187-218.

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, Harafuji N, Hastings KE, Ho I, Hotta K, Huang W, Kawashima T, Lemaire P, Martinez D, Meinertzhagen IA, Necula S, Nonaka M, Putnam N, Rash S, Saiga H, Satake M, Terry A, Yamada L, Wang HG, Awazu S, Azumi K, Boore J, Branno M, Chin-Bow S, DeSantis R, Doyle S, Francino P, Keys DN, Haga S, Hayashi H, Hino K, Imai KS, Inaba K, Kano S, Kobayashi K, Kobayashi M, Lee BI, Makabe KW, Manohar C, Matassi G, Medina M, Mochizuki Y, Mount S, Morishita T, Miura S, Nakayama A, Nishizaka S,

Nomoto H, Ohta F, Oishi K, Rigoutsos I, Sano M, Sasaki A, Sasakura Y, Shoguchi E, Shin-i T, Spagnuolo A, Stainier D, Suzuki MM, Tassy O, Takatori N, Tokuoka M, Yagi K, Yoshizaki F, Wada S, Zhang C, Hyatt PD, Larimer F, Detter C, Doggett N, Glavina T, Hawkins T, Richardson P, Lucas S, Kohara Y, Levine M, Satoh N, Rokhsar DS. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science. 298:2157-67

Doak TG, Witherspoon D, Doerder FP, Williams KR, Herrick G. 1997. Conserved features of ciliate TBE1 transposons. Genetica 101:75-86.

Doak TG, Doerder FP, Jahn C, Herrick G. 1994. A family of transposases genes in transposons found in prokaryotes, multicellular eukaryotes and ciliated protozoans. Proc Natl Acad. Sci USA 91:942-946.

Doak TG, Witherspoon DJ, Jahn CL, Herrick G. 2003 Selection on the genes of *Euplotes crassus* Tec1 and Tec2 transposons: evolutionary appearance of a programmed frameshift in a Tec2 gene encoding a tyrosine family site-specific recombinase. Eukaryot Cell. 2:95-102.

Doak TG, Cavalcanti AR, Stover NA, Dunn DM, Weiss R, Herrick G, Landweber LF. 2003 Sequencing the *Oxytricha trifallax* macronuclear genome: a pilot project. Trends Genet. 19:603-7.

Dos Ramos F, Carrasco M, Doyle T, Brierley I. 2004 Programmed -1 ribosomal frameshifting in the SARS coronavirus. Biochem Soc Trans. 32:1081-3.

Eddy SR. 2005 A model of the statistical power of comparative genome sequence analysis. PLoS Biol. 3:e10.

Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, Badger JH, Ren Q, Amedeo P, Jones KM, Tallon LJ, Delcher AL, Salzberg SL, Silva JC, Haas BJ, Majoros WH, Farzad M, Carlton JM, Smith RK Jr, Garg J, Pearlman RE, Karrer KM, Sun L, Manning G, Elde NC, Turkewitz AP, Asai DJ, Wilkes DE, Wang Y, Cai H, Collins K, Stewart BA, Lee SR, Wilamowska K, Weinberg Z, Ruzzo WL, Wloga D, Gaertig J, Frankel J, Tsao CC, Gorovsky MA, Keeling PJ, Waller RF, Patron NJ, Cherry JM, Stover NA, Krieger CJ, del Toro C, Ryder HF, Williamson SC, Barbeau RA, Hamilton EP, Orias E. 2006 Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol. 4:e286.

Erbeznik M, Yao MC, Jahn CL. 1999. Characterization of the *Euplotes crassus* macronuclear rDNA and its potential as a DNA transformation vehicle. J Euk. Microbiol. 46:206-216.

Foissner, W., H Berger. 1999. Identification and Ontogenesis of the nomen nudum Hypotrichs (Protozoa: Ciliophora) Oxytricha nova (= Sterkiella nova sp. n.) and O. trifallax (= S. histriomuscorum). Acta Protozool.38: 215 - 248

Fouche N, Moon IK, Keppler BR, Griffith JD, Jarstfer MB. 2006 Electron microscopic visualization of telomerase from *Euplotes aediculatus* bound to a model telomere DNA. Biochemistry. 45:9624-31.

Gall JG. 1981. Chromosome structure and the C-value paradox. J Cell Biol 91:3-14.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B. 2002 Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature. 419:498-511.

Gottschling DE, Cech TR. 1984. Chromatin structure of the molecular ends of *Oxytricha* macronuclear DNA: phased nucleosomes and a telomeric complex. Cell 38:501-510.

Gottschling DE, Zakian VA. 1986. Telomere proteins: specific recognition and protection of the natural termini of *Oxytricha* macronuclear DNA. Cell 47:195-205.

Herrick G. 1994. Germline-soma relationships in ciliated protozoa: the inception and evolution of nuclear dimorphism in one-celled animals. Sem Dev Biol 5:3-12.

Herrick G, Wesley RD. 1978. Isolation and characterization of a highly repetitious inverted terminal repeat sequence from *Oxytricha* macronuclear DNA. Proc. Nat. Acad. Sci. USA 75:2626-2630.

Hoffman DC, Prescott DM. 1997 Evolution of internal eliminated segments and scrambling in the micronuclear gene encoding DNA polymerase alpha in two *Oxytricha* species. Nucleic Acids Res. 25:1883-9.

Hoffman DC, Anderson RC, DuBois ML, Prescott DM. 1995. Macronuclear gene-sized molecules of hypotrichs. Nucleic Acids Res 23:1279-1283.

Klobutcher LA. 2005 Sequencing of random *Euplotes crassus* macronuclear genes supports a high frequency of +1 translational frameshifting. Eukaryot Cell. 4:2098-105.

Klobutcher LA, Herrick G. 1997. Developmental genome reorganization in ciliated protozoa: the transposon link. Prog Nucleic Acid Res and Mol Biol 56:1-62.

Klobutcher LA, Swanton MT, Donini P, Prescott DM. 1981. All gene-sized DNA molecules in four species of hypotrichs have the same terminal sequence and an unusual 3' terminus. Proc Natl Acad Sci USA 78:3015-3019.

Jahn CL, Klobutcher LA. 2002. Genome remodeling in ciliated protozoa. Ann. Rev. Microbiol. 56:489-520.

Juranek SA, Rupprecht S, Postberg J, Lipps HJ. 2005 snRNA and heterochromatin formation are involved in DNA excision during macronuclear development in stichotrichous ciliates. Eukaryot Cell. 4:1934-41.

Kumar S, Filipski A. 2007 Multiple sequence alignment: in pursuit of homologous DNA positions. Genome Res. 17:127-35.

Landweber LF, Kuo TC, Curtis EA. 2000. Evolution and assembly of an extremely scrambled gene. Proc Natl Acad Sci USA. 97:3298-30.

Liang H, Wong JY, Bao Q, Cavalcanti AR, Landweber LF. 2005 Decoding the decoding region: analysis of eukaryotic release factor (eRF1) stop codon-binding residues. J Mol Evol. 60:337-44.

Linger J, Cech TR. 1996. Purification of telomerase from *Euplotes aediculatus*: Requirement of a primer 3' overhang. Proc Natl Acad Sci USA 93:10712-10717

Lingner J, Hughes TR, Shevchenko A, Mann M, Lundblad V, Cech TR. 1997 Reverse transcriptase motifs in the catalytic subunit of telomerase. Science. 276:561-7.

Lozupone CA, Knight RD, Landweber LF. 2001. The molecular basis of nuclear genetic code change in ciliates. Curr Biol. 11:65-74.

Maercker C, Kortwig H, Lipps HJ. 1999. Separation of micronuclear DNA of *Stylonychia lemnae* by pulsed-field electrophoresis and identification of a DNA molecule with a high copy number. Genome Res 9:654-61.

Meyer E, Garnier O. 2002. Non-Mendelian inheritance and homology-dependent effects in ciliates. Adv. Genet. 46:305-337.

Miceli C, Ballarini P, Di Giuseppe G, Valbonesi A, Luporini P. 1994 Identification of the tubulin gene family and sequence determination of one beta-tubulin gene in a cold-poikilotherm protozoan, the antarctic ciliate *Euplotes focardii*. J Eukaryot Microbiol. 41:420-7.

Mochizuki K, Gorovsky MA. 2005 A Dicer-like protein in *Tetrahymena* has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. Genes Dev. 19:77-89.

Möllenbeck M, Postberg J, Paeschke K, Rossbach M, Jonsson F, Lipps HJ. 2003. The telomerase-associated protein p43 is involved in anchoring telomerase in the nucleus. J Cell Sci. 116:1757-61.

Möllenbeck, M., A. R. O. Cavalcanti, F. Jönsson, H. J. Lipps, and L. F. Landweber. 2006. Interconversion of germline-limited and somatic DNA in a scrambled gene. J. Mol. Evol. 63(1):69-73.

Namy O, Rousset JP, Napthine S, Brierley I. 2004 Reprogrammed genetic decoding in cellular gene expression. Mol Cell. 13:157-68.

Nanney DL. 1981. T. M. Sonneborn: an interpretation. Annu Rev Genet. 15:1-9.

Neafsey DE, Hartl DL, Berriman M. 2005 Evolution of noncoding and silent coding sites in the *Plasmodium falciparum* and Plasmodium reichenowi genomes. Mol Biol Evol. 22:1621-6.

Orias E. 2000. Toward sequencing the *Tetrahymena* genome: exploiting the gift of nuclear dimorphism. J Eukaryot Microbiol 47:328-333.

Paschka AG, Jönsson F, Maier V, Möllenbeck M, Paeschke K, Postberg J, Rupprecht S, and Lipps HJ. 2003. The use of RNAi to analyze gene function in spirotichous ciliates. Eur. J. Protistol. 39:449-454.[

Pollard, D.A., Bergman, C.M., Stoye, J., Celniker, S.E., and Eisen, M.B. 2004. Benchmarking tools for the alignment of functional noncoding DNA. BMC Bioinformatics 5:6.

Pluta AF, Kaine BP, Spear BB. 1982. The terminal organization of macronuclear DNA in *Oxytricha fallax*. NAR 10:8145-54.

Riley JL, Katz LA. 2001. Widespread distribution of extensive chromosomal fragmentation in ciliates. Mol Biol Evol 18:1372-1377.

Rubin GM et ~49 other authors. 2000. Comparative genomics of the eukaryotes. Science 287:2204-2215.

Seegmiller A., Williams KR, Hammersmith RL, Doak TG, Messick T, Witherspoon D, Storjohann LL, Herrick G. 1996. Internal eliminated sequences interrupting the *Oxytricha* 81 locus: allelic divergence, conservation, conversions, and possible transposon origins. Mol Biol Evol 13:1351-1362.

Soldo AT, Brickson SA, Larin F. 1981. The kinetic and analytical complexities of the DNA genomes of certain marine and fresh-water ciliates. J. Protozool 28:377-383.

Sonneborn TM. 1975. *Tetrahymena pyriformis*. in Handbook of Genetics Vol 2 (RC King ed) Plenum Press, NY, pp 433-467 (Plenum, NY).

Sonneborn,T. M., 1957 Breeding systems, reproductive methods and species problems in Protozoa. pp. 155-329. In: The Species Problem. Edited by E. Mayr. AAAS, Washington, D.C. - ,

Sonneborn TM. 1977. Genetics of cellular differentiation: stable nuclear differentiation in eucaryotic unicells. Annu Rev Genet 11:349-367.

Steinbruck G, Haas I, Hellmer KH, Ammermann D. 1981 Characterization of macronuclear DNA in five species of ciliates. Chromosoma. 83:199-208.

Stone EA, Cooper GM, Sidow A. 2005. Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. Annu Rev Genomics Hum Genet. 6:143-64

Subramanian S, Kumar S. 2003 Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res. 13:838-44.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigó R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES. Initial sequencing and comparative analysis of the mouse genome. 2002. Nature. 420(6915): 520-62

Wright ADG, Lynn DH. 1997. Maximum ages of ciliate lineages estimated using a small subunit rRNA molecular clock: Crown eukaryotes date back to the Paleoproterozoic. Archiv. Protistenkd. 148: 329-341.

Yao M-C, Duharcourt S, Chalker D. 2002. Genome-wide rearrangements of DNA in ciliates. In: Mobile DNA II. Craig NL, Craigie R, Gellert M, Lambowitz AM, eds. ASM Press, Washington, D.C., pp. 730-758.