Sequencing the *Tetrahymena thermophila* Genome

A White Paper

Submitted to the National Human Genome Research Institute

February 10, 2002

Submitted by:

Eduardo Orias
Research Professor of Genomics
Coordinator of the *Tetrahymena* Genome Sequencing Project
Department of Molecular, Cellular and Developmental Biology
University of California, Santa Barbara,
Santa Barbara, CA 93106
Phone: (805) 893 3024
Fax (805) 893 4724
orias@lifesci.ucsb.edu

In consultation with:

The Whitehead Institute Center for Genome Research

Overview: *Tetrahymena* as a valuable *genetic unicellular animal* model

Tetrahymena thermophila belongs to the Alveolates, a major evolutionary branch of eukaryotic protists composed of three primary lineages: Ciliates (e.g., *Tetrahymena* and *Paramecium*), Dinoflagellates (e.g., *Symbiodinium*, the coral endosymbiont, and *Alexandrium*, which causes paralytic shellfish poisoning) and the exclusively parasitic Apicomplexa (e.g., *Plasmodium falciparum*, the causative agent of malaria*)*. *Tetrahymena thermophila* is a ciliated protozoan belonging to a free-living, fresh-water genus that is highly successful ecologically. No *free-living* alveolate genome has been sequenced.

Since 1923, when Nobel Laureate Andre Lwoff succeeded in growing *Tetrahymena* in pure culture, two sibling species of the genus *Tetrahymena* (*pyriformis* and *thermophila*) have been used as *microbial animal models.* With the development of genetic methods in *T. thermophila* in the 1950's, this has become the species of choice throughout the field.

Tetrahymena has typical eukaryotic biology. Its ultrastructure, cell physiology, development, biochemistry, genetics, and molecular biology have been extensively investigated. This organism displays a degree of cellular structural and functional complexity comparable to that of human and other metazoan cells. Consistent with this, analyses of mRNA complexity and very recent EST projects have confirmed that, at the molecular level, *Tetrahymena*'s rich and complex genome conserves a rich set of ancestral eukaryotic functions [1]. In addition, *Tetrahymena*'s special elaborations of certain basic eukaryotic mechanisms have facilitated discoveries opening the door to major new fields of fundamental research, including:

- First cell whose division was synchronized, leading to the first insights into the existence of cell cycle control mechanisms.
- Identification and purification of the first cytoskeletal motor, dynein, and determination of directional activity.
- Participation in the discovery of lysosomes and peroxisomes.
- One of earliest molecular descriptions of programmed somatic genome rearrangement.
- Discovery of the molecular structure of telomeres, telomerase enzyme, the templating role of telomerase RNA and their roles in cellular senescence and chromosome healing.
- Nobel-prize winning co-discovery of catalytic RNA (ribozymes);
- Discovery of the function of histone acetylation in transcription.

The richness of *Tetrahymena*'s biology makes it a genetic unicellular animal model organism "for all seasons." An impressive array of novel molecular genetic technologies places *Tetrahymena* at the forefront of experimental, *in vivo* functional genomics research [2], and complements a wealth of favorable biological features. Sustained extramural grant support of *Tetrahymena* research and published statements by leading researchers working on other organisms attest to the importance of *Tetrahymena's* contributions [3-7]. Availability of the *Tetrahymena* genome sequence will have major benefits in molecular bioscience and biotechnology. Areas of impact include 1) fundamental biological and biomedical research; 2) finding the function of predicted human genes with homologs in *Tetrahymena* but not in yeast; 3) value for experimental functional genomics and 4) informing the biology of other alveolates, including pathogens of major medical or agricultural significance.

This white paper, which responds to specific encouragement from the Trans-NIH NonMammalian Models Committee, seeks the completion of whole-genome shotgun-sequencing and at least partial closure of the *Tetrahymena* macronuclear (MAC) genome. It is submitted on behalf of the *Tetrahymena* research community, in consultation with the Whitehead Institute Center for Genome Research. This white paper will be a) distributed through the ciliate molecular biology list server (supervised by Prof. Jacek Gaertig at the University of Georgia); b) placed in the *Tetrahymena* genome website, http://www.lifesci.ucsb.edu/~genome/Tetrahymena**,** and be available for downloading by FTP from http://www.lifesci.ucsb.edu/~orias/ftp. Recent advances in molecular genetic tools for functional genomics in *Tetrahymena*, described in this paper, have been highlighted in a recent review [2].

A. Specific biological rationales for the utility of new sequence data

Genome sequence-enabled comparative genomics has become a major stimulus for hypothesis-driven research in modern biomedical science. The richness of its genome and its key phylogenetic position make *Tetrahymena* an important model organism for this purpose. Ultimately however, definitive biological mechanistic understanding is gained only by experiment. This places a premium on model organisms with facile genetic and molecular tools that allow the use of the genomic sequence for experimental analysis. *Tetrahymena* has recently emerged as an outstanding example of these rare organisms. The rest of this white paper develops this theme and responds in detail to the NHGRI questionnaire.

1. Improving human health.

Tetrahymena is an excellent model system for finding the functions of human genes. A high fraction of *Tetrahymena* ESTs match human proteins, many of which have no homologs in yeast, the benchmark unicellular eukaryotic genetic model organism. Given the ~30,000 genes estimated to exist in the *Tetrahymena* genome, we expect that thousands of *Tetrahymena* proteins will have homology with important human proteins not represented in *S. cerevisiae.* Furthermore, humans share a high degree of *functional conservation* with ciliates. This is evidenced by better matches of *Tetrahymena* EST [1] and *Paramecium* coding sequences [8] to humans than to non-ciliate microbial genetic model organisms.

Sequence similarity conserved over more than a billion years of independent evolution of humans and *Tetrahymena* predicts a) that the function of the genes is important in both organisms -- and thus likely to cause human hereditary disease by dysfunctional mutation -- and b) that the proteins have likely retained their basic, ancestral biochemistry. Thousands of human genes of unknown function are predicted by analysis of the human genome sequence. Sequence conservation is a valuable criterion for prioritizing which ones to study. The combination of genome richness, sequence conservation, favorable biological features and powerful molecular genetic tools, should confer on the biomedical research community an enormous opportunity to use *Tetrahymena* experimentally to obtain a better understanding of the molecular basis of many diseases, and for improving human health.

2. Informing human biology.

Tetrahymena is a well-established model organism for the study of fundamental molecular, cellular and developmental biology. Major areas currently under active investigation include 1) cell motility, 2) developmentally programmed DNA rearrangements, 3) regulated secretion, 4) phagocytosis, function of post-translational modifications of 5) tubulins and 6) histones and 7) telomere maintenance and function. The first five areas represent important human biology that *cannot* be investigated in *S. cerevisiae*.

Fundamental research in *Tetrahymena* has developed advanced molecular genetic tools (see Section B2b) and has established productive paradigms of post-genomic experimental analysis. In research areas where conserved protein components have already been identified in *Tetrahymena*, the tools for postgenomic analysis have quickly led to recent important discoveries. Such areas include the essential functions of post-translational phosphorylation of histone in transcription (initiated in *Tetrahymena*) and of post-translational polyglycylation of tubulin in maintenance of axoneme stability and sensitivity of longitudinal cytoskeletal microtubules to cell-cycle-controlled severing. These tools and experimental paradigms, *in combination with the genome sequence*, should profoundly stimulate discovery in other important areas of fundamental investigation:

- Research that would immediately benefit from identifying *Tetrahymena* homologs of proteins implicated by work in other organisms:
    - Telomere structure, telomerase enzymology, and their cellular regulation

- Chromosome replication and copy number maintenance
- Cytoskeletal motors
- Cytoskeleton function and regulation, cytoskeletal specialization
- Phagocytosis and phagosome-mediated bacterial pathogenesis
- Regulated apoptosis
- Chemoreception and signal transduction

- Research that would immediately benefit from large-scale cell fractionation and high throughput mass spec analysis, coupled with high quality genomic sequence:
  - Determination of the complete complement of ciliary proteins, phagosome proteins and proteins involved in the regulated secretion of protein storage granules
  - Characterization of microtubule functional diversity: *Tetrahymena* has 17 distinct microtubule systems including ciliary axonemes, centriolar structures and mitotic spindles

- Research that would immediately benefit from high throughput mRNA expression profiling analysis:
  - Developmentally regulated, immunoglobulin-gene-like chromosome breakage-rejoining (chromatin diminution)
  - Developmentally regulated gene amplification
  - Germline and soma differentiation and maintenance
  - Determination of the basis for mating type determination, sexual maturation and senescence

3. Informing the human *sequence*.
- *Tetrahymena* can inform features of the human sequence by the investigation of the function of many *ab initio* predicted genes, as described earlier.
- Functional RNA genes may be more readily predicted in *Tetrahymena* due to the high AT content of noncoding sequences.
- *Tetrahymena* possesses unique biological advantages for the study of ribosomal RNA synthesis, processing and function: a) a *single* germline copy of the 18S and 28S rRNA genes; b) homogeneous, small (21 kb) MAC chromosome exclusively dedicated to those rRNA genes and maintained at 9,000 copies per cell; c) Many nucleoli (~500 per MAC) that are purifiable. Thus availability of the *Tetrahymena* sequence, in combination with mass spec approaches and the advanced genetic tools, has the potential to allow a full understanding of nucleolar biology.

4. Providing a better connection between the sequences of non-human organisms and the human sequence.

Tetrahymena is a well-studied genetic unicellular animal model. Experimental investigations at the molecular and cell level are easier in *Tetrahymena* than in metazoans because of rapid growth rate and clonal homogeneity of cell cultures. Furthermore, some of the biology shared by humans and *Tetrahymena* is missing not only in yeast but even in the invertebrate metazoan genetic model organisms (*Drosophila* and *C. elegans*). Examples are specialized paralogs of the tubulin gene family (delta, epsilon, eta) found in ciliary basal bodies. These structures are close homologs of centrioles, which function in human mitotic division. Thus, investigations of functions of predicted human genes in *Tetrahymena* would *complement* and facilitate their investigation at more integrative levels, i.e., using *multicellular* animal models.

5. Expanding our understanding of basic biological processes relevant to human health.
Many observations suggest the potential benefits of the *Tetrahymena* genome sequence for investigating human neurobiology. *Tetrahymena* has opioid receptors with pharmacological properties similar to human ones, and is already being used as a model to test natural marine compounds that inhibit pain and inflammation. *Tetrahymena* EST or GSS (genome survey sequence) reads match receptor components for two other brain neurotransmitters, GABA and NMDA. *Tetrahymena* cells also possess catecholamines. A handful of ESTs match KIAA predicted proteins, sequenced from mRNAs expressed

in the human brain, some of which are absent in yeast. A *Tetrahymena* GSS sequence, recently obtained at TIGR, matches a transmembrane protein expressed in the mouse cochlea, whose mutation causes deafness. This preview, based on a miniscule sample of sequence reads, illustrates the likely abundance of important health-related genes whose function can be studied by molecular genetic methods in *Tetrahymena*.

Telomerase has been implicated in human tumorigenesis and cellular aging, and has become a major biomedical research area. Greater understanding of telomere structure, telomerase enzymology, and their cellular regulation would be very useful, and *Tetrahymena* is an excellent model organism for these investigations. There is greater similarity between human and *Tetrahymena* telomerases, and likely telomeres, than between human and budding yeast or other model organisms. Furthermore, telomerase has been efficiently reconstituted from purified components in vitro *only* by using *Tetrahymena* components. A *Tetrahymena* gene database would quickly enable the identification and experimental investigation of homologs of relevant proteins identified in other organisms. Such studies could facilitate the development of better therapeutics for human diseases of telomerase insufficiency (somatic cell proliferative deficiencies) and hyperactivation (cancer).

Phagocytosis is an important and conserved but poorly understood cellular process. Since the phagosome is the primary route of invasion of many microbial pathogens, a better understanding of its biology should also lead to novel strategies for fighting pathogen invasion and improving human health. *Tetrahymena* phagosomes can be purified in much larger scale than mouse macrophage phagosomes. Availability of the *Tetrahymena* genome sequence would allow determination of the full protein composition of a conserved eukaryotic phagosome, enabling identification and experimental analysis of the function of mammalian homologs.

The *Tetrahymena* sequence should also reveal genetic functions missing in parasitic alveolates (e.g., malaria parasite) that are likely supplied by the human host. Such information might be of help in developing strategies to combat parasites and protect human health.

6. Providing additional surrogate systems for human experimentation

Tetrahymena is an excellent surrogate model for animal research. The promise for discovering the cellular and molecular basis of many diseases has been described earlier. This work would render unnecessary much preliminary research in animals.

Tetrahymena also has an enormous potential for drug testing, made possible by functional similarity to human cells, fast growth, clonally-homogeneous cell culture and readily visualized and quantifiable physiological endpoints. These include growth rate, phagocytosis rate, induced exocytosis, swimming speed and direction, chemotaxis, osmoregulation (contractile vacuole pulse rate), cytokinesis, conjugation, meiosis induction, and nuclear differentiation. In addition, *Tetrahymena* has hundreds of cilia. They provide large amount of plasma membrane for the high level expression of surface proteins, which are high priority targets for drug development by the pharmaceutical industry. For example, surface proteins with vaccine potential from two parasitic protists, the malarial parasite *Plasmodium* and the fish ciliate parasite "Ich" (Ichthyophthirius), have already been expressed in the plasma membrane of *Tetrahymena*. The likely existence of homologs of many brain neurotransmitters receptors is another area where the *Tetrahymena* genome sequence could have an important impact as a surrogate animal system, e.g., in the study of analgesic and anti-inflammatory compounds already underway.

Tetrahymena is a favorite organism for toxicological tests and for the study of quantitative structure/activity relationships (QSAR) among environmental toxicants. A database of *Tetrahymena* QSARs for more than 2000 compounds is available [9]. Environmental toxicity assays are important for the protection of human health and *Tetrahymena's* advantages allow it to be used as an inexpensive surrogate for fish-based lethality tests.

7. Facilitating the ability to do experiments in *Tetrahymena*.
Some of the benefits that would accrue from the genome sequence have already been noted under section A2. In addition, *Tetrahymena* has superior tools for sequence-enabled experimental analysis by "reverse genetics", i.e., going from gene sequence to mutant phenotype (see Section B2b). The genome sequence will also facilitate "forward genetics", i.e., from mutant phenotype to gene sequence.
- *Tetrahymena* is a genetic model organism with a well-developed facility for forward genetics (see section B2b), using methods suitable for high-throughput analysis.
- For mutant phenotypes accompanied by growth selection, among other methods, cloning by complementation has become feasible using whole-genome DNA and the highly inducible metallothionein promoter [10].
- For mutations not accompanied by growth selection, genetic coassortment analysis facilitates positional mapping by narrowing down gene location to within a single MAC chromosome [18], or, on the average, to within 100 genes. Availability of the sequence will then allow the identification of the gene by DNA-mediated recombination rescue with mixture of cloned inserts or PCR products from the relevant MAC chromosome.

8. Expanding our understanding of evolutionary processes.
The alveolates offer one-to-two-billion years of deep eukaryotic evolution and diversity. Representing the first *self-standing* genome from the alveolate clade, the *Tetrahymena* genome sequence will provide robust information on the full complement of genes of early eukaryotes. Examples of specific potential sequence-enabled contributions from this highly complex unicell are highlighted below.

Evolution of the role of positional information in cell architecture and development. Ciliate cells (including *Tetrahymena*) maintain orthogonal axes of polarity that specify analogs of cellular longitude and latitude. At binary fission, these gradients provide precise coordinates for the development and positioning of highly differentiated, unique cortical structures for daughter cells, such as the oral apparatus (the site of phagosome formation), the cytoproct (the site of phagosomal egestion) and the contractile vacuole (the site of active water expulsion). *Tetrahymena* mutations that disrupt these developmental gradients have been extensively analyzed at the cellular level. The genome sequence, and associated tools for forward and reverse genetics, would greatly accelerate analyses of the molecular bases for these phenomena, providing valuable insights into the evolution of metazoan development.

Evolution of germline vs. soma differentiation. The ciliates are an experiment of nature in which germline vs. soma differentiation (silent micronucleus vs. expressed macronucleus) is restricted to the nuclear apparatus of *a single cell*. Germline vs. soma differentiation, prevalent in the metazoan and higher plants, is nearly unique to the ciliates in the eukaryotic protist world. In addition, *Tetrahymena* has at least seven newly discovered members of the piwi/argonaute gene family, which functions in stem cell maintenance in metazoa and plants. These are being actively investigated and the first one has been shown to be essential for development of the somatic macronucleus [Mochizuki, Fine, Gorovsky and Pearlman, pers. comm.]. Availability of the *Tetrahymena* genome sequence should make additional important contributions to the understanding of the evolution of such fundamental developmental processes as germline/soma differentiation.

The *Tetrahymena* genome sequence can also contribute valuable insights in other areas, including evolution of the genetic code (UAR, along with CAR, are glutamine codons in *Tetrahymena*) and evolution of immunoglobulin-like DNA rearrangements that occur during MAC differentiation.

B. Strategic issues in acquiring new sequence data

1. The demand for the new sequence data.
The ciliate research community currently includes more than 300 active molecular and cell biologists in ~150 research groups; the majority works with *Tetrahymena*. Additional ciliate investigators work on areas of ecology and evolution. To our knowledge, the ciliate community is the *largest using a genetic model organism without a genome project*. The community is cross-linked by a web of scientific

collaboration, sabbaticals, visits to laboratories to learn new techniques and coauthorships of published articles. The highly collaborative nature of the community has amplified ideas and resources -- and thus its productivity and the quality of its contributions -- well beyond what might be expected from a sheer body count. The community publishes currently ~300 papers per year -- 338 are listed in PUBMED for 2001. Sequence-enabled stimulating discoveries should lead to the expansion of the *Tetrahymena* community. Prospective postdocs, university positions and granting agency support are likely to gravitate toward an excellent genetic model organism *that has a genome project* and that grows rapidly and cheaply and requires relatively little special expertise or equipment to use. Furthermore, the sequence will put the very complete animal proteome and the advanced experimental tools of *Tetrahymena* in the service of fundamental, biomedical and applied research by the *general scientific community*. The collaborative research stimulated by this development is an additional factor that should lead to further expansion of the community that uses *Tetrahymena* for biomedical research.

The *Tetrahymena* community has a high degree of enthusiasm for this genome-sequencing project, because of the eagerly anticipated acceleration of on-going research in many areas of fundamental significance where the cutting edge work is being done on *Tetrahymena*. The genome sequence is also anticipated by members of the protist research community. Enthusiastic letters of support, included with the NIGMS application (see Section B5), were received from Thomas Cech (U. of Colorado, Boulder and President of HHMI, Nobel Laureate); C. David Allis (University of Virginia, NIH Stetten Lecturer); Michael Gray (Dalhousie University, Director of the Protist EST Project); Kathleen Collins (U. of California, Berkeley); Joseph Frankel (U. of Iowa); Martin Gorovsky (U. of Rochester); Patrick Keeling (U. British Columbia); Laura Landweber (Princeton University); Ronald Pearlman (York University, Toronto); and Linda Sperling (CNRS, Gif-sur-Yvette, France). In addition, 28 committed ciliate biologists contributed important materials for the concept paper submitted to the Trans-NIH NonMammalian Models Committee in November 2001, the precursor of this white paper.

Tetrahymena community involvement with the genome project started in August 1999 at a *Tetrahymena* Genomics Workshop held in conjunction with the (biennial) International Conference on Ciliate Molecular Biology. A second *Tetrahymena* Genomics Workshop was organized at the next Ciliate Molecular Biology meeting in July 2001. Both workshops were plenary sessions, attended by the majority of the participants. Plans and important issues were circulated to the community in advance of both conferences, ensuring wide-ranging discussion and facilitating consensus about the genome project. The first workshop resulted in the formation of a Steering Committee for the *Tetrahymena* Genome Project, and the selection of E. Orias as the project coordinator. The Committee, which consists of 18 internationally recognized molecular biologists in the ciliate research community, has since met yearly, and has interacted extensively by email and phone calls. Reports of Steering Committee meetings and the concept paper submitted to the Trans-NIH NonMammalian Models Committee (which can be downloaded from http://www.lifesci.ucsb.edu/~orias/ftp) were circulated to the entire ciliate community.

2. Suitability of *Tetrahymena* for experimentation.

a) Favorable biological features [11].

Fastest growing microbial animal model (as short as 1.5 hr doubling time). Unicellularity and fast growth enable culture homogeneity, quick experimental results, low maintenance costs, and compact space requirements.

Dual, self-sufficient nutritional modes: particle (bacteria) phagocytosis and small-molecule uptake by active transport. Cells can be grown axenically (in pure culture) and in chemically defined medium. This allows complete control of the chemical and physical growth environment.

Large cell size (50 x 30 micrometers): facile injection, cytology, immunocytology and FISH, electrophysiological recording.

Clonal growth to high density under wide volume range (microdrops to bioreactors).

Facile large-scale cell fractionation.

Large temperature range for growth (18^{O}C-41^{O}C), giving great latitude in experimental conditions.

Simple freezing protocol allows long term maintenance and germline protection of valuable strains in liquid nitrogen.

Abundance of species in genus with well-characterized phylogeny, including close and distant relatives: useful for decryption of regulatory DNA elements and functional RNA domains and for evolutionary studies.

Well-defined life cycle including sexual cycle (conjugation), sexual immaturity, sexual maturity and sexual senescent stages of vegetative growth.

Mitosis and meiosis restricted to a germline nucleus that is not essential for growth, and gene-specific transcription restricted to a non-mitotic nucleus: facilitates *independent* experimental analyses of these fundamentally important processes.

Developmentally-regulated apoptosis of parental macronucleus during new macronuclear differentiation.

b) Tractability for genetic studies.

A recent volume of Methods in Cell Biology [12] reviews the well-developed *Tetrahymena* genetics and contains detailed protocols. We highlight below unusual and powerful genetic approaches enabled by biological features that accompany germline/soma differentiation [see also 2; 13].

Readily inducible self-fertilization, leading in a single step to *whole-genome micronuclear (MIC) and macronuclear (MAC) homozygotes*.

Heterokaryons, i.e., cell lines in which the MIC and MAC differ genotypically, are readily constructed and are used in myriad applications. They are especially useful in positive selection of conjugant progeny, and for the facile maintenance of lethal mutations, aneuploidy (chromosome losses or gains) and essential gene knockouts in the *homozygous state* in the silent germline of heterokaryons.

Allelic assortment in the macronucleus allows independent genetic mapping of loci to MAC chromosomes, wide range of wild type to mutant allele ratios, and direct DNA-mediated transformation of the MAC.

DNA-mediated transformation, routinely accomplished by electroporation, biolistic bombardment or microinjection; mass transformation rates $>10^4$ transformants per microgram of DNA. Transforming DNA can be selectively targeted to the MIC, the differentiating MAC or the mature MAC. Integrative, high-copy-replicative and developmental processing vectors are each available.

Genomic integration of linearized recombinant DNA occurs *exclusively by precise homologous recombination*, allowing highly specific gene replacement, disruption (knockout) and targeted insertion of foreign genes. The high specificity of insertion also enables the targeted characterization of individual members of a family of very similar genes.

High frequency of co-transformation allows each of two constructs to specifically integrate at their own, separate homologous locus, allowing efficient indirect selection for a desired replacement.

Gene over-expression is obtained by high-copy-number vectors, allowing 200-fold gene amplification relative to the rest of the genome, and/or an inducible metallothionein (MTT) promoter allowing experimentally controlled regulation of gene expression over a 1000-fold dynamic range.

Ribosomal antisense repression by a novel and robust approach in which e*very* ribosome in the cell displays the same antisense sequence of the targeted gene. Antisense repression is stably maintained and clonally inherited. Gene-specific ribosomal antisense mutagenesis is a novel extension of ribosomal antisense repression that allows efficient phenotype-based cloning of mutant genes from an antisense library ("forward genetics").

Mutants with defined phenotypes, affected in major cell processes (e.g., ciliary motility, phagocytosis, regulated secretion, cytokinesis, developmental positional information, chromosome stability, etc.). Living up to a long tradition, all mutants are freely available to anyone from the *Tetrahymena* genetic stock collection and individual laboratories, even prior to publication.

3. Rationale for obtaining the complete genomic sequence of *Tetrahymena*.

The richness of the *Tetrahymena* proteome and the availability of powerful tools for post-genomic experimental analysis can be exploited to greatest advantage by sequencing *the entire genome*. To obtain the high quality sequence of every gene, the only equivalent alternative would be cDNA sequencing. But all cDNA-based strategies are unavoidably incomplete because they are very sensitive to the level of gene expression and developmental regulation. Furthermore, cDNAs do not provide enough flanking sequence information to facilitate gene replacement constructs.

In many other eukaryotes (e.g., human and other metazoans) gene finding is facilitated by ESTs or cDNA sequencing because coding sequence is a small fraction of the genomic sequence and is interrupted by sizeable introns. That advantage is much less important in *Tetrahymena* where a) introns are relatively rare and small; they occupy at most 35% of the transcribed genome; b) protein-coding sequences are readily identified, as they differ markedly in A+T composition (~62%) from the rest of the MAC genome sequence (~83%), i.e., introns, intergenic and subtelomeric regions; and c) there is only a single termination codon (UGA) and a reasonably well-conserved sequence surrounding the initiator AUG; d) no alternative splicing has been reported [14; 15]. Thus sequencing the entire macronuclear genome is a more time and cost efficient way to obtain the entire set of genes than any additional EST or cDNA sequencing effort. Funding already available will allow the sequencing of tens of thousands of *Tetrahymena* EST within the next year, which will provide a statistically reliable data set for training HMM-model-based gene finding programs.

Tetrahymena possesses a silent, germline (micronuclear) genome and an expressed (macronuclear) genome. An alternative to macronuclear sequencing would be to sequence the germline (micronuclear) genome. Given funding limitations, we have opted to sequence the expressed (macronuclear) genome because it retains all the genes and other DNA elements required for the life of the organism, while eliminating most of the repeated sequence and selfish DNA elements present in the germline genome [20].

4. The cost of sequencing the genome and the state of readiness of the organism's DNA for sequencing.

The macronuclear genome is estimated to have ~180 Mb, or less than 6% of a mammalian genome. It is estimated to contain 20-40,000 genes, comparable to the number of genes in the human genome and ~5 times larger than that in the yeast genome. There is no evidence of significant genome or gene duplications. Preparations of *Tetrahymena* wild type macronuclear DNA, purified by the method of Gorovsky et al. [16], are available. When additional DNA is needed, hundreds of micrograms can be prepared and tested for purity within one week after culture inoculation.

It would be most desirable to obtain *genome sequence that is as completely finished as possible.* This would multiply the benefits of the *Tetrahymena* genomic sequences for the following independent, scientifically important reasons:

Experimental *in vivo* functional genomics: one of the most valuable sequence-enabled experimental tools available in *Tetrahymena* is gene replacement/knockout by exact homologous recombination. Efficient replacement requires hundreds of bp of flanking homologous sequence. The unrestricted ability to do gene replacements and knockouts with high throughput technology would be guaranteed only by having finished sequence of essentially the entire intergenic regions, given the high coding density.

Proteomics: *Tetrahymena* presents an enormous opportunity in the field of functional proteomics. Its metazoan-like cellular complexity occurs within a single large cell, amenable to large-scale fractionation, starting from physiologically homogeneous clonal cultures. The entire set of components of many important organelles could be identified and opened to functional investigation by proteomic, e.g. mass spec, analyses. Only finished gene sequence can guarantee the success of such analyses.

Phylogenetic: *Tetrahymena* would be the first free-living representative of the entire Alveolate clade to have a genome sequence. Finished genome sequence should facilitate investigations of the biology, not just of other ciliate model organisms, but also of a variety of alveolates of medical and agricultural importance.

Developmental chromosome diminution and germline/soma evolution: studies of the immunoglobulin-like internal deletions and of germline/soma evolution will require knowledge of the germline (MIC) sequence. The high throughput mapping and identification of MIC-limited segments, which occur outside coding sequences, should be facilitated by comparisons of MIC WGS of limited sequence coverage with finished intergenic MAC sequence.

The following strategy will be used to sequence the *Tetrahymena* macronuclear genome:

a) Whole-genome shotgun (WGS) sequencing. The WI-CGR has experience with whole-genome-shotgun sequencing and assembly for genomes ranging from 5.8 Mb to 2.7 Gb. The challenge presented by *Tetrahymena's* high AT composition will be addressed by obtaining deep sequence coverage from small-insert clones (that exhibit high stability) and long-range links to tie these sequence contigs together. Specifically, we will generate at least 10x whole-genome shotgun sequence in paired-end reads from 4-kb plasmids (90%) and from jumping libraries (10%; see below). This corresponds to a total of approximately 4.2 M attempted reads, assuming a pass rate of 80% and an average Phred 20 read length of 540 b. This sequence will provide a combined physical coverage of approximately 80x (~45x with 4-kb plasmids, ~35x with jumping libraries, derived from 40-kb fragments) that should ensure the generation of a high-quality assembly.

The task is quite feasible: (i) the total number of reads needed corresponds to just over one month of WI-CGR's current capacity, and (ii) the WI-CGR routinely constructs high-quality WGS plasmid and fosmid libraries from randomly sheared DNA. Since inserts larger than 6 kb seem to be unstable, presumably due to the AT-rich sequence of *Tetrahymena* (average 75%), the WI-CGR would also construct a 40-kb jumping library to provide the long links necessary to achieve large sequence scaffolds. The inserts in this library would be smaller than 6 kb and consist of sequence from both ends of randomly sheared 40-kb fragments (WI-CGR routinely prepares these fragments for construction of fosmid libraries). The feasibility of this approach has been demonstrated by the sequencing and assembly of whole Plasmodium chromosomes, which have even a higher AT content (>80%), both at TIGR and the Sanger Center.

b) Closure of the genomic sequence. We believe that producing a high quality, deep shotgun assembly should be the highest priority for this project. In addition, we recognize the value in providing finished sequence to the user community. However, until we assemble our shotgun sequences, cloning bias and thus the actual number and size of the gaps in the *Tetrahymena* genome cannot be assessed. Closure of the genome regions that have the highest AT content may present a challenge because we may lack clones to serve as sequencing templates. As a first step to improve the quality of the *Tetrahymena* assembly WI-CGR will perform one round of automated prefinishing (= large-scale transposon tagging) to close gaps that are spanned by plasmid clones. The extent to which finishing should be carried out can be prioritized later, on the basis of evolving assessment of cost and capacity.

c) Access to the *Tetrahymena* genome data. Genome data will be released in accordance with NHGRI rules. All traces will be submitted to the NCBI trace archive. The *Tetrahymena* community is organizing a *Tetrahymena* Genome Database, to be hosted by the Saccharomyces Genome Database at Stanford University. The WI-CGR has an ongoing collaboration with the Stanford group to distribute and display sequence and associated genomic information. The *Tetrahymena* community will work closely with WI-CGR and Stanford to make the data available in a form that will be maximally used.

The above plan was developed in close consultation with WI-CGR. Discussions on *Tetrahymena* genome-sequencing started in August 2001 and culminated in December 2001, when E. Orias met at WI-CGR with Bruce Birren and James Galagan to discuss concrete sequencing plans and gave an invited talk on special features and advanced genetic tools available in *Tetrahymena*. Additional discussions with James Galagan took place at a December 2001 TIGR workshop on Prospects for Protist Genomics, where E. Orias delivered a talk on *Tetrahymena* genomics. Final consultation with Bruce Birren and Nicole Stange-Thomann, by e-mail and phone, took place this past week.

A diversity of genomic resources, already available in *Tetrahymena*, will significantly facilitate various phases of the sequencing project.

Genetic maps. Germline linkage maps and macronuclear coassortment maps have been constructed based on ~400 DNA polymorphisms and linking an estimated 2/3 of the genome [17; 18; http://www.lifesci.ucsb.edu/~genome/Tetrahymena]. In addition, more than 100 partial deletions of germline chromosomes have been mapped [Cassidy-Hanley et al., unpubl. obs.].

Physical maps. Micronuclear sequence reads flanking an estimated 15% of the chromosome breakage sites (Cbs) have been obtained and they have all been assigned to MIC chromosome arms. The

physical size of the two MAC chromosomes pairs flanking each of those Cbs has also been determined [Hamilton, Cassidy-Hanley et al., submitted]. Funding is available to characterize the rest of the Cbs (estimated at ~ 300 total) This sequence data will eventually coassemble with MAC genome sequence and will allow determination of the order and orientation of MAC chromosomes in the germline genome.

About a third of the Cbs junctions have DNA polymorphisms, which have been mapped to linkage groups. In addition, more than 60 genetically mapped RAPD polymorphisms located on distinct MAC chromosomes have been sequenced and the size of the MAC chromosome has been determined. Another ~60 are being done (Orias lab). This work will anchor physical and sequence maps to the genetic map and facilitate the long-range assembly of genomic sequence. Funding is also available to the Orias lab to physically map, over the next three years, at least 2,000 sequenced-tagged sites (mainly ESTs), which will ultimately anchor the sequence map to the physical map of the genome.

The availability of anchored genetic, physical and sequence maps will not only facilitate the overall assembly of the genome sequence, but should also facilitate positional cloning of mutant genes whose phenotype does not provide selective growth advantage in either direction.

Proteins and ESTs. More than 150 experimentally characterized and annotated genes from *T. thermophila* (mainly) and *T. pyriformis* have been deposited in GenBank -- some are genomic, others are mRNA sequences. About 500 non-redundant ESTs, derived from full-length cDNA library from exponentially growing cells [19], have been sequenced and submitted (or about to be submitted) to GenBank [1; http://www.cbr.nrc.ca/reith/tetra/tetra.html]. Funding is also available to sequence an additional 20-40,000 ESTs from several libraries, mainly through a subproject, under Prof. Ron Pearlman's direction, of the Protist EST Project of the Atlantic Division of Genome Canada. The ESTs will be useful, not only for gene discovery, but also for training *Tetrahymena* gene finding programs.

5. Other sources of funding available or being sought for this sequencing project

Available: A $70,000 fund for EST sequencing, led by Aaron Turkewitz (University of Chicago), has been built by seed funds awarded by his university and supplemented by the contributions of members of the *Tetrahymena* research community --including private funds of some members. In addition, the Genome Canada Initiative (Atlantic Division) has funded the Protist EST Project (PEP), which includes a budget for the sequencing of 20-40,000 *T. thermophila* ESTs, as indicated above.

Being sought: NIGMS agreed to accept an R01 application for WGS sequencing of the *Tetrahymena* genome, which was submitted for the February 1, 2002 deadline and will follow the normal NIH review process. This project, a collaborative effort of TIGR, the *Tetrahymena* research community and the Saccharomyces Genome Database, has three aims: 1) To whole-genome shotgun sequence and assemble the macronuclear genome to a depth of 8-fold sequence coverage over a 3-year period. 2) To electronically annotate and analyze the genome sequence, including the identification of putative genes, prediction of gene function, and other features standard for genomic analysis. 3) To facilitate unrestricted, user-friendly access to the *T. thermophila* genome sequence by releasing the sequence data immediately to external sequence databases and by the creation of three interlinked database resources: a TIGR website, a manually curated *Tetrahymena* Genome Database and a *Tetrahymena*-specific section in the NCBI "Genomic Biology" website. We understand that even if the project is funded, however, NIGMS may not be able to support the full 8x sequence coverage requested. Additional funding sources to supplement the proposed NIGMS project are being actively sought, by applying to a joint NSF/USDA initiative to sequence microbial genomes, with a May 1, 2002 application deadline, and by seeking a contribution from Genome Canada.

Selected References

1. Fillingham, J., N. Chilcoat, A. Turkewitz, E. Orias, M. Reith, and R. Pearlman, *Analysis of expressed sequence tags (ESTs) in the ciliated protozoan Tetrahymena thermophila.* J. Euk. Microbiol., 2002. In press. A preprint can be privately downloaded from http://www.lifesci.ucsb.edu/~orias/ftp (two files).
2. Turkewitz AP, Orias E & Kapler G (2002) Functional Genomics: The coming of age for *Tetrahymena thermophila*. Trends in Genetics, 18:35-40.
3. Lundblad V (1998) Telomerase catalysis: a phylogenetically conserved reverse transcriptase. Proceedings of the National Academy of Sciences of the United States of America, 95:8415-16.
4. Rosenbaum J (2000) Cytoskeleton: functions for tubulin modifications at last. Curr. Biol., 10:R801-3.
5. Gull K (2001) Protist tubulins: new arrivals, evolutionary relationships and insights to cytoskeletal function. Curr. Opin. Microbiol., 4:427-32.
6. Hutton JC (1997) *Tetrahymena*: the key to the genetic analysis of the regulated pathway of polypeptide secretion? Proceedings of the National Academy of Sciences of the United States of America, 94:10490-92.
7. Kirschner M, Gerhart J & Mitchison T (2000) Molecular "vitalism". Cell, 100:79-88.
8. Dessen P, Zagulski M, Gromadka R, Plattner H, Kissmehl R, Meyer E, Betermier M, Schultz JE, Linder JU, Pearlman RE, Kung C, Forney J, Satir BH, Van Houten JL, Keller AM, Froissard M, Sperling L & Cohen J (2001) *Paramecium* genome survey: a pilot project. Trends in Genetics, 17:306-8.
9. Schultz TW (1997) TETRATOX: The *Tetrahymena pyriformis* population growth impairment endpoint - A surrogate for fish lethality. Toxicol. Meth., 7:289-309.
10. Shang Y, Song X, Bowen J, Corstanje R, Gao Y, Gaertig J. & Gorovsky MA (2002) Proc. Nat. Acad. Sci., in press.
11. Orias E, Hamilton EP & Orias JD (1999) *Tetrahymena* as a laboratory organism: Useful strains, cell culture and cell line maintenance. In: Asai DJ & Forney JD (eds.), *Tetrahymena thermophila*. Meth. Cell Biol. Academic Press, New York, NY, p. 187-209.
12. Asai DJ & Forney JD (1999) *Tetrahymena thermophila*. Methods in Cell Biology. Academic Press, San Diego. 580 p.
13. Karrer KM (1999) *Tetrahymena* genetics: Two nuclei are better than one. In: Asai DJ & Forney JD (eds.), *Tetrahymena thermophila*. Meth. Cell Biol. Academic Press, New York, NY, p. 127-86.
14. 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. Journal of Biological Chemistry, 258:6899-6905.
15. 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., 46:239-47.
16. Gorovsky MA, Yao MC, Keevert JB & Pleger GL (1975) Isolation of micro- and macronuclei of *Tetrahymena pyriformis*. Methods. Cell. Biol., 9:311-27.
17. Wickert S & Orias E (2000) *Tetrahymena* micronuclear genome mapping: A high resolution map of chromosome 1L. Genetics, 154:1141-53.
18. Wickert S, Nangle L, Shevel S & Orias E (2000) *Tetrahymena* macronuclear genome mapping: colinearity of macronuclear coassortment groups and the micronuclear map on chromosome 1l. Genetics, 154:1155-67.
19. Chilcoat ND, Elde NC & Turkewitz AP (2001) An antisense approach to phenotype-based gene cloning in *Tetrahymena*. Proceedings of the National Academy of Sciences of the United States of America, 98:8709-13.
20. Yao MC & Gorovsky MA (1974) Comparison of the sequences of macro- and micronuclear DNA of *Tetrahymena pyriformis*. Chromosoma, 48:1-18.