PROJECT SUMMARY

The overall goals of this proposal are to obtain a high-coverage, high-quality draft of the *Acanthamoeba castellanii* Neff (*Ac)* genome and annotate and analyze the genome sequence data. We intend to use the whole genome shotgun method followed by whole genome assembly to construct a high-quality genome assembly for *Ac*. The project team is well suited to undertake this work. The PI, Brendan Loftus is a faculty member at The Institute for Genomic Research (TIGR) and PI on the *Entamoeba histolytica* genome effort amongst other eukaryotic projects, and has publications in genome assembly and analysis. The Co-PI, Michael Gray, is Canada Research Chair in Genomics and Genome Evolution at Dalhousie University in Halifax, Nova Scotia, Canada. Dr. Gray also holds an appointment as Fellow in the Program in Evolutionary Biology, Canadian Institute for Advanced Research. Dr. Gray has extensive experience in protist mitochondrial genomics, is currently Project Leader of the Protist EST Program (PEP), a large-scale genomics initiative managed and partly funded by Genome Canada and has specific responsibility for several of the individual PEP EST projects, including that for *Ac*.

The intellectual merit of the proposal stems from the significant position of *Ac* as one of the few well-studied members of the free-living amoebae, which play an important role in a variety of terrestrial and aquatic environments. As such, *Ac* is one of the most commonly found amoebae in soil. From an evolutionary perspective *Ac* is proposed to be a member of the ancient subphylum Protamoebae. Knowledge of its genome sequence will improve not only the understanding of the origins of the Phylum Amoebozoa but also provide a suitable outgroup for understanding of some of the special features of the other amoebae whose genomes are currently being sequenced, including *Dictyostelium* and the *Entamoeba histolytica*. *Ac* has also been shown to harbor a wide variety of symbionts, a number of which are important intracellular pathogens and bio-defense organisms. It is postulated that *Ac* not only acts as an environmental reservoir, facilitating survival and dispersal of these organisms, but also induces phenotypic changes leading to increased invasiveness and virulence capabilities in mammalian hosts.

The broader impacts of this study are that it will allow for the application of functional genomics approaches towards a better understanding of the amoebic lifestyle, adaptation to low oxygen environments and the balance between the choice of a predator-prey and symbiont relationship between bacteria and amoebae. It will permit a better understanding of lateral gene transfer between ameboid hosts and their symbionts and provide an important resource in the use of *Ac* as a model system for studying certain intracellular pathogens and their amoeboid hosts. The genome resources should be of interest to a wide variety of scientific disciplines including soil ecologists, environmental microbiologists, medical microbiologists, scientists studying cell structure and motility in amoebae, protistologists, evolutionary biologists and others studying various aspects of this organism as a human pathogen. The data provided should foster a variety of cross-disciplinary studies including comparison of the various pathogenic and non-pathogenic amoebae, comparison of the relationship between intracellular pathogens with both their protozoan and mammalian hosts, and a broader understanding of the many indirect but important contributions of protozoa to a variety of environments

PROJECT DESCRIPTION

A. SPECIFIC AIMS

Aim 1: Obtain a high coverage, high-quality draft of the *Acanthamoeba castellanii* Neff (*Ac)* genome and annotate and analyze the genome sequence.

Aim 2: Attempt to identify *Ac* genes that may be informative for future studies on motility, the encystations process, nitrogen fixation and survival in hypoxic environments.

Aim 3: Use of Comparative genome analysis using data from other ongoing amoebic genome projects (*Dictyostelium discoideum*, (*Dd*) and *Entamoeba histolytica (Eh)*) to help determine the genetic component of the amoebic lifestyle.

Aim 4: The use of a phylogenomics approach to identify potentially laterally transferred genes within the Ac genome and compare these data with lateral gene transfer candidates being identified from other completed and ongoing protist genome projects.

Aim 5: Identify genes that many be involved in uptake and support of a wide range of symbionts including a variety of intracellular pathogens.

B. BACKGROUND AND SIGNIFICANCE

1. Acanthamoeba castellanii as a key well studied member of the small free-living amoebae and an important constituent of a variety of ecosystems

Bacterial communities are central to the functioning of a wide variety of terrestrial communities and are heavily grazed by bacteriophagous microfauna including amoebae. This grazing affects the community structures by reducing overall numbers of bacteria and stimulating the

mineralization of nutrients. This phenomenon has been poorly studied in soil communities; however a number of investigations have shown that protozoan grazing has substantial effect on the taxonomic and morphological makeup of bacteria in marine environments (50) (42). *Acanthamoeba castellanii* is an important member of the small free-living amoebae (FLA) which act as a link between macrocosms and microcosms in soils where they are the main predators of bacteria (33) (36). These ubiquitous small FLA feed on many kinds of microbes, including bacteria, fungi, yeasts, algae, and other amoebae. Small FLA populate the boundary zones between water and other media such as soil (highest abundance), air, plants and animals, where their food sources are also concentrated, and there is a critical balance between their numbers and those of the organisms on which they subsist. Within soils, pore size controls the concentration and distribution of microorganisms and because of their small size FLA are capable of widespread penetration within these pores. In the soil, bacteria and their various predators, including amoebae, are found in pores of various sizes, and of particular interest, in the rhizosphere, on and near plant roots (41). Since many of the rhizosphere bacteria play a major role in root growth and nitrogen fixation, by grazing on these bacteria, FLA, together with other protozoa, have a major effect on plant growth, mineralization and the nutrient cycle through increases in the levels of nitrogen and phosphorous recycling (9, 41). Additionally, *Acanthamoeba* and other free-living amoebae themselves have enzymes that can metabolize nitrogen (48). The most common species of soil protozoa include *Hartmannella*, *Naegleria* as well as *Acanthamoeba* sp. *Acanthamoeba* sp are also found in aquatic habitats which is an environment normally too harsh for most soil protozoa (14). In general *Acanthamoeba* sp are more tolerant of extremes besides being one of the most common of the genera derived from the atmosphere (37). Ac has also found comfortable niches in urbanized environments including the atmospheres of cities, urban water supplies and sewage systems (45). Bacteria have been shown to be more resistant to chlorine treatment following ingestion by, and survival within protozoa. This bacterium-protozoan association can lead to persistence of bacteria in chlorine-treated water and may be one of the mechanisms which have evolved as a mechanism for survival of fastidious bacteria in dilute and inhospitable aquatic environments (23).

2. Significance of the Acanthamoeba castellanii genome sequence for fundamental biology.

Investigating poorly understood mechanisms of fundamental significance.

Major areas currently under active investigation in Acanthamoeba include the following 1) understanding the amebic lifestyle 2) Ac as an important component of the soil ecosystem 3) understanding the mechanisms of the symbiont versus predator-prey relationship between Ac and different bacteria 4) understanding Acanthamoeba castellanii's role as an environmental reservoir and amplification vehicle for a variety of intracellular pathogens. 5) unusual metabolism in Ac and 6) the potential for, and the effects of, lateral gene transfer on the genome of Ac.

Understanding aspects of and gaining insights into the amoebic lifestyle

Ac has been a better studied organism than many of the other free living amoebae and there is an important growing body of literature on its molecular interactions. In particular, Ac has served as a model organism for studies on the cytoskeleton, cell movement, and aspects of signal transduction and gene regulation (28) (20). Generation of a high quality assembly of the Ac genome would allow its sequence and genome architecture to be compared with genome data

from other amoeboid organisms. Ac is phylogenetically distant from the other amoeboid organisms whose genome projects are currently underway (43) and investigation of its gene content should permit the identification of genes that are common and essential to the amoebic lifestyle, but vary in molecular components that differentiate them. This is currently difficult comparing data solely from the various Entamoeba species and the free-living amoeba *Dictyostelium discoideum* as the parasitic lifestyle of *E. histolytica* and its relatives has altered dramatically the constituents of their genomes.
Given that EST data are known to be incomplete and represent a randomly selected collection of expressed information that will vary from organism to organism, reliance on EST sequence data alone is generally insufficient for such comparisons, especially if the relevant genes are expressed at a low level. The Ac genome sequence would represent the only other free-living amoebic representative and would undoubtedly assist in the interrogation and interpretation of the genome data from the model organism *Dictyostelium discoideum*. Ac also a pathogen and comparison of the three genome datasets should also reveal common gene sets between *Ac* and *Eh* and absent From *Dd* which may represent elements required for amoebic pathogenicity. Many other well-known amoebae (*e.g., Amoeba proteus*) have extremely large genomes that are clearly not appropriate sequencing targets with the currently available sequencing technology.

Ac as important component of the soil ecosystem

Ac is one of the predominant soil organisms in terms of population size and distribution and an important inhabitant of water communities. Knowledge of its protein repertoire would thus be of interest in reconstructing its metabolism to help determine its role in the nitrogen and phosphorus cycles, mineralization, and other aspects of its ecosystem. Unlike the ciliates, most flagellates, the rotifera and other large amoeba (*e.g., Amoeba proteus*), Ac can penetrate small pores in the soil and is thus more likely to occupy broader swathes of the rhizosphere. As a consequence it is likely than these organisms to play a larger role in geochemistry than other protists with a more limited ecological range.

Acanthamoeba castellanii as host to a variety of symbionts and a model system for the study of the relationships between amoebae and their symbionts

For reasons that are poorly understood, *Acanthamoeba,* more so than other soil amoebae, is often likely to harbor endosymbiotic bacteria (39). Ac is capable of supporting the growth of bacteria such as *Parachlamydia acanthamoeba* (18), *Salmonella typhimurium* (17), *Escherichia coli* O157 (4), *Legionella pneumophila* (40), *Francisella tularensis* (1), *Burkholderia pseudomallei* (21), *Helicobacter pylori* (51), *Coxiella burnetii* (26), *Chlamydia* (31) (13), *Pseudomonas aeruginosa* (8), *Mycobacterium avium* (32), *Stenotrophomonas maltophilia* and *Serratia marcescens* (47)], the soil fungus *Cryptococcus neoformans* (44), and yeast (2). This reinforces Ac's role in a variety of environments not only as regulators of the overall numbers of various bacterial and other species but as an important environmental reservoir for the maintenance of a variety of bacterial symbionts. It is reasonable to assume that a subset of Ac's symbionts are obligate symbionts while for others the symbiont relationship is a facultative one. There is no good understanding as to the molecular basis of obligate versus facultative symbiosis or why certain bacteria have the ability to destroy or be destroyed by their amoebic hosts. It has been shown that there are differences in the ability of Ac grown in axenic culture to lyse different

bacteria: *Bacillus megaterium*, *Bacillus subtilis*, *Chromatium vinosum*, *Micrococcus luteus* and *Pseudomonas fluorescens* are easily lysed under these conditions, whereas *Agrobacterium tumefaciens*, *Klebsiella aerogenes* and *Serratia marcescens* are quite refractory (49). To better understand what determines how bacteria and amoebae choose between a predator-prey and a commensal relationship, it is important to understand the molecular biology and genetic framework of each organism. While a few hundred bacteria (including a number of symbionts) have been or are in the process of being sequenced, few amoebic genomes have been or are being sequenced. The complex interaction between amoebic hosts and their symbionts is also evidenced by the fact that the presence of a symbiont has an effect on gene expression of certain genes within amoebae (22). Additionally it appears that obligate bacterial endosymbionts are able to enhance amoebic pathogenic potential in vitro by some as-yet unknown mechanism (16).

Acanthamoeba castellanii as an amplification vehicle and model system for the study of a variety of intracellular pathogens.

A wide variety of intracellular pathogens have now been shown to reside and be able to survive within Acanthamoeba including *Escherichia coli* O157 (4), *Legionella pneumophila* (40), *Francisella tularensis* (1) and other bacteria and fungi as noted above. Acanthamoebae appear to play at least two critical roles in the dissemination of these intracellular pathogens, firstly by acting as environmental reservoirs whereby growth and survival within protozoa can provide protection to other microbes from harsh or inhospitable environments. Because of the ubiquitous nature of Acanthamoebae this capacity can facilitate dissemination and spreading of the organism through wind-borne transmission of amoebic cysts or through human and animal ingestion of airborne or water-borne amoebae. Secondly, phenotypic changes as a result of intra-protozoal growth not only lead to enhanced environmental fitness but also to increased invasion and virulence in mammalian hosts. Study of this apparent co-evolution of intracellular pathogenic bacteria and their lower-order predators will enhance our understanding of many environmental pathogens. *Acanthamoeba castellanii* because of its significant and widespread presence in various environments provides a unique opportunity to study such interactions. Studies of the pathogenicity of *Legionella pneumophila* and *Coxiella burnetii* have already been carried out using Ac as a model system (52).

Unusual metabolism of Ac

Ac is able to survive under very low oxygen and reducing conditions (e.g., in deeper layers of underwater sediments). In order to survive under these conditions it likely requires alternative electron acceptors, although none have been identified for this organism. *E. histolytica* is obligately microaerophilic, and it recycles electrons from glycolysis through fermentation to form ethanol, acetate, and carbon dioxide (30). Several of the enzymes involved in fermentation in *Eh* have been acquired by lateral gene transfer from anaerobic prokaryotes (34). *Ac* has not been reported to ferment glucose, and its adaptations to low oxygen environments likely differ from those of *E. histolytica*. It will be interesting to contrast how the two amoebae survive under these conditions, and to see if Ac has adapted to low oxygen by acquiring genes from prokaryotes. *D. discoideum*, on the other hand, is strictly aerobic and is very sensitive to hypoxic conditions (3). It will be interesting to contrast the genomes of these three amoebae in terms of their ability or lack of ability to live under low oxygen conditions.

Lateral gene transfer

The enormous recent interest and focus on the potential for, and the consequences of, lateral (or horizontal) gene transfer (LGT, or HGT) has up until recently focused on gene transfer between prokaryotes and more recently from prokaryotes to eukaryotes (19) (25). Preliminary data from the *E. histolytica* genome project indicates that LGT appears to be a significant feature in the evolution of this genome. Emerging data from genome projects from two other amitochondriate protists (*Giardia* and *Trichomonas*) appears to point to similar findings as do EST data for Ac (see preliminary data). However many of the current examples that supports extensive potential LGT events come from parasitic protozoa which makes it difficult to determine whether these observations are likely to be general feature of protist genomes or restricted to members of certain protist lifestyles. Aside from determining the absolute numbers, it would be interesting to determine whether or not the same classes of genes appear to be candidates for LGT in Ac considering the known presence of symbionts within this organism. Inclusion of these data might well have significant implications for the generation and interpretation of the various theories on the source and the direction of LGT within eukaryotes. An example of the potential consequences of such events are evidenced by the fact that that LGT has been proposed to account for the observed occurrence of plant like genes in the human pathogen Chlamydia trachomatis (http://www.ncbi.nlm.nih.gov/Coffeebreak/CB1_Chlamydia/page.html#). While more speculative, it is of interest that *L. pneumophila*, known passengers within Ac, have homologs of the *Agrobacterium tumefasciens* Ti plasmid vir genes (40), and in both bacteria, these Type IV secretion system genes can transfer macromolecules from bacteria to host cells (amoebae and plants respectively); if Ac also phagocytizes A. tumefasciens, they may serve as exchange vessels for this postulated LTS event. Cataloging the genes followed by phylogenetic analysis of both host and symbiont in these cases, allows more detailed inferences on the occurrences and consequences of these events in major components of the biosphere.

Understanding Microbial diversity

The taxonomy of the Protista is extraordinarily complex. One approach to understanding this complexity has been simply to place protist lineages into 71 groups without identifiable sister taxa, along with 200 additional lineages lacking clear identities; one estimate is that there are 200,000 named species (15). Traditionally, four broad categories of protists have been recognized: flagellates, amoebae, algae and parasitic protists, the latter group including members of all the other groups. In one scheme ("Tree of Life"; http://tolweb.org/tree/eukaryotes/accessory/amoebae.html), the amoebae, traditionally defined by the presence of pseudopods with which they move and feed, are assigned to the taxon "Sarcodina", which includes 20 groups of heterotrophic amoebae and 59 additional named species with no clear ultrastructural identity. A. castellanii is a key member of the protozoan phylum Amoebozoa. Recent phylogenetic analyses position the Amoebozoa as the closest protozoan group to the ophisthokonts (animals + fungi), within one of two major eukarotic groups, the "Unikonts" (primitively uniciliate eukaryotes) (43). Because the root of the eukaryotic tree appears to lie between unikonts and bikonts (biciliated eukaryotes), and because the Amoebozoa are the earliest diverging unikonts known, comparison of Amoebozoan genome sequences will be invaluable in inferring the genetic information complement of the unikont common ancestor, as well as the last common ancestor of eukaryotes as a whole. Currently there are a number of ongoing major amebic genome projects including *Eh* and *Dd* both of which are closer to each other than to *Acanthamoeba castellanii*. *A. castellanii* is proposed to be a member

of the ancient subphylum Protamoebae and will improve not only the understanding of the origins of the Amoebozoa phyla but also provide a suitable out-group for understanding of some of the special features of the other sequenced amoebae including the muticellular fruiting phase of *Dictyostelium* and the adaptations to parasitism in the various Entamoeba species (see attached letters of support).

Favorable features of Acanthamoeba castellanii for post genomic experimental analysis.

The generation of a high-quality draft of the *Ac* genome would provide a wealth of data for the generation of testable hypotheses. However, additional benefits would be gained by the generation of tools for functional genomic studies, and determination of the genome sequence of *Ac* would clearly pave the way for a number of functional studies. Firstly it would allow for comparison of any gene-based (using comparative genome hybridization) and transcriptional differences between symbiont-containing strains of *Ac* compared to axenically-grown strains, as well as pathogenic and non-pathogenic strains (see attached letters of support). The *Ac* genome sequence should also pave the way for functional studies of the strategies for survival of the various symbionts, particularly intracellular pathogens within *Ac*, many of whose genomes have been sequenced and for whom DNA micro-arrays are available or are becoming available. In contrast there do not appear to be any symbionts in the pathogenic amoeba *E. histolytica*. Thus, it will be of interest to study the phago-lysosome gene complements of both soil amoebae, *Dd* and *Ac*, compared with those of *Eh*. Such studies can be pursued through DNA-based microarray analysis. The relatively small size of the *Ac* genome with a consequently small number of genes should make the generation of such microarrays a feasible goal. Finally, newer methodologies such as RNA interference (RNAi) that are being tried successfully on other protests including Paramecium and with other amoebae including *Dd* and *Eh* are likely to prove to be usable in *Ac*.

The *A. castellanii* genome and its suitability for sequence-enabled genetic analysis

DNA renaturation kinetics, in combination with information on the quantity of DNA per cell and mitochondrial:nuclear DNA ratios, has been used to derive an approximate unique genome size of 33 Mb for *A. castellanii* (7). The GC content is expected to be close to 60% based on buoyant density measurements (7). This represents a reasonable and tractable genome size and GC composition for whole genome shotgun sequencing and assembly.

When considering generation the level of genome coverage a balance needs to be struck between the desire to generate not only the sequence of virtually all of the genes within the genome but also to recreate the architecture of the genome while avoiding the most costly aspects of finishing or closure. The finishing phase is time consuming, expensive and requires substantial informatics support. Because of the highly repetitive content of both amebic genomes currently underway, generation of a high degree of coverage using a variety of insert sized plasmids will be necessary to achieve the clone coverage required to achieve a stable accurate genome assembly. Generation of a stable genome assembly is important when attempting to determine gene structures in *Ac* many of which are thought to contain introns. In addition a stable assembly containing long range sequence scaffolds allows for easier detection of presence or absence of particular genes, determination of orthologs, differentiation between orthologs and paralogs, comparison of syntenic regions between organisms and studies of genome evolution. We

propose to generate 8-fold sequence coverage of the genome using a wide variety of plasmid sizes. Possible problems that may arise during the sequencing and assembly phase include polymorphisms within the chosen strain; mitochondrial contamination, the possibility of aneuploidy across certain regions of the genome, and the possible presence of bacterial symbionts within the Neff strain. We have attempted to control for some of these potential problems by using highly purified (using a CsCl gradient) nuclear DNA designed to remove mitochondrial DNA cross-contamination and using a presumed axenic strain to reduce the possibility that it is contaminated by endosymbiotic bacterial DNA. The preliminary results cited below seem to indicate that mitochondrial or symbiont contamination will not be major issues in this effort. However we plan on an ongoing basis to keep track of polymorphic positions through continuous genome assembly and analysis to detect the presence of aneuploidy, detect and remove mitochondrial contamination by filtering with the already completed mitochondrial genome sequence. We will also use ongoing and completed microbial genome data publicly available to help determine on an ongoing basis the presence of contaminating sequences from symbionts and other sources.

C. PRELIMINARY STUDIES/PROGRESS REPORT

1. Available *A. castellanii* genomic resources

Genome-wide libraries and preliminary genome and EST sequence data.

We have generated two individual 6-8 kb libraries from *Ac* nuclear DNA cloned into the pHOS2 vector (described below). Approximately 3750 good sequences were generated from each library to give a final number of 7506 good sequences. The average success rate of the sequences was 89% and the average edited length of the sequences was 810 bp indicating that the organism appears to sequence well. Assembly of the data using the Celera Assembler gave 5724 singletons and 792 assemblies for an overall of 5.68 Mb of unique sequence and 0.2 X coverage indicating the representative nature and the randomness of the libraries. Running Repeatmasker using the default settings gave no obvious repeats using a library designed for use with *Eh*. In order to detect mitochondrial contamination of the library the assembled sequences were searched against the published mitochondrial sequence for Ac (accession no: NC_001637). Only 11 sequences gave significant hits to the mitochondrial sequence indicating that the DNA used to make the libraries appears essentially free of mitochondrial contamination and indicates that such contamination would not be a significant issue in a whole genome shotgun effort. In order to determine the level of any symbiont or other bacterial contamination we searched the assembled sequences against a dataset of 254 publicly available completed and ongoing bacterial and archeal genome projects. Only 28 of the sequences gave significant blast matches at the nucleotide level and generally the matches were short indicating that the random sequences generated appear to be essentially free of known bacterial contaminants. This indicates that the problem of contaminating symbiont DNA would not be a significant issue in a whole genome effort. We also searched the assembled sequences against the *A. castellanii* singleton EST dataset provided by Dr Gray. In all 778 of the assembled sequences matched the ESTs at identities greater than 97% nucleotide sequence identity. They matched 2002 EST singletons indicating a redundancy in the EST dataset and the merit of a whole genome shotgun approach. Similarly searches of a database of predicted proteins gave 2616 matches with an expect score of < 1 e-5 indicating the usefulness of the approach for gene discovery. Using gapped alignment methods

to align the EST sequences with the genome assemblies it is apparent that Ac has a significant complexity within its gene structures with lots of exons being identified and one transcript containing at least 9 exons being identified. Overall these data indicate that the genome is tractable to sequence and that the genomic DNA can be effectively cleaned of contaminants.

Analysis of data from the *A. castellanii* EST project to determine similarity to data from the ongoing genome projects of Eh and Dd.

When searching the Ac EST data against all available open reading frames (ORFs) from Dd using tblastx greater than 61% of the Ac assembled EST contigs show no blast hit, 16% show a blast score (expect value) of between e-05 and e-20 and only ~22% of the assembled ESTs show a blast score of < e-20. When searched against all available ORFs from the Eh genome project, 43% show no significant blast hit, 20% show a score of between e-05 and e-20 and only ~36% show scores of < e-20. The ORF data from both Eh and Dd represents data from almost complete coverage of both genomes. This data demonstrates the large degree of sequence divergence between Ac and both Eh and Dd indicating that many novel genes are likely to emerge from an Ac genome sequence and also that any conserved regions between the organisms are likely to be significant in understanding the genic basis of the amoebic lifestyle.

Gene Index.

An *A. castellanii* EST project has recently been launched, again using this particular strain, under the auspices of the "Protist EST Program" (PEP), an initiative partially funded by Genome Canada. 12,000 ESTs representing >6000 unique clusters are available through PEPdb, the database of the Protist EST Program. These will be used to generate a Gene Index for *Ac* similar to those created for other protists (http://www.tigr.org/tdb/tgi/protist.shtml) which will be created by March 2004.

***2.* Ongoing eukaryotic genome projects at TIGR**

TIGR is now WGS sequencing to varying degrees of completion a number of eukaryotic genomes, including the Apicomplexans *Theileria parva* (9 Mb, 4 chromosomes) and *Plasmodium yoelii* (25Mb, 14 chromosomes), the fungi *Aspergillus fumigatus* (30 Mb), *Coccidoides immitis* (28 Mb), and *Cryptococcus neoformans* (19 Mb) and the amoebae *E. histolytica (20 Mb). T. parva* has been completed and is awaiting publication. *P. yoelii* has been sequenced to 5X coverage and used in a comparative study with *Plasmodium falciparum.* The fungi *A. fumigatus* and *C. neoformans* have been completed and are in the publication preparation stages. Other large scale genome projects currently underway include *Tetrahymena thermophila* (180 Mb) which has been sequenced to 8X coverage, *Trichomonas vaginalis* (~75 Mb) which has been sequenced to 5X, *Brugia malayi* (110 Mb) which has been sequenced to 3X. A more complete and up to date status for these and other organisms can be obtained from (http://www.tigr.org/tdb/euk/). As a result of these and other sequencing projects, TIGR has developed the infrastructure and methods required for rapid, cost-effective shotgun sequencing of intermediately sized genomes such as *Acanthamoeba castellanii.* Significant and necessary advances in quality and throughput have been achieved in each of the following areas: 1) DNA library construction, 2) template production; 3) DNA sequencing and quality assessment; 4) genome assembly; 5) quality control. These advances are summarized in the following sections:

Library construction. The construction of high quality random plasmid libraries is critical for the success of a shotgun -sequencing project. Central to this success is constructing libraries with a relatively tight insert size range, few to no clones without inserts, and no clones with chimeric inserts. These three elements are controlled by nebulization and size selection of insert DNA and a robust library construction protocol. We have constructed a series of vectors (pHOS) containing BstXI cloning sites. These vectors include several features: 1) the sequencing primer sites immediately flank the BstXI cloning site to avoid excessive re-sequencing of vector DNA; 2) the vectors are propagated in *E. coli* at different copy numbers; pHOS2 used for constructing 3-4 kb and 10 kb shotgun libraries has a copy number of about 25, and pHOS3 used for constructing the largest shotgun libraries has a copy number of one (BAC replicon); *3*) the elimination of strong promoters oriented toward the cloning site, thus minimizing the possibility of insert-coded toxic peptide synthesis and insert transcription-stimulated recombination events. TIGR is also using linking libraries for WGS sequencing projects. These linking libraries have only the ends of 50 kb clones contained in the HOS2 vector, separated by a Kan^R cassette. This library is prepared from sheared genomic DNA - 50 kb fragments. They are ligated into the pHOS 2 vector and the resultant constructs are cut with a restriction enzyme that does not cut in the vector and leaves on the average between 2 and 4 kb of DNA from each end ligated to the vector. A kanamycin-resistance cassette is then ligated into the construct and the clones are selected with kanamycin and ampicillin after electroporation into *E. coli*. These linking libraries are more stable than traditional large insert libraries since they actually only contain ~4-8 kb of DNA from the target genome. The 50 kb linking library strategy has been successfully employed at TIGR in projects on eukaryotes e.g., *A. fumigatus, C. neoformans, B malayi, T. vaginalis* and others.

Template Preparation and Sequencing Reactions. Libraries in the form of ligation mixes are stored in multiple aliquots at 4oC. Records for each library are generated and stored in a tracking database at TIGR. When required for high-throughput sequencing, a library is transformed, and cells are plated onto large format (16cm x 16cm) diffusion plates, prepared by layering 150ml of fresh molten agar lacking antibiotic onto a previously set 50ml layer of agar containing antibiotic. Cells are plated as soon as the top layer solidifies, allowing time for the antibiotic resistance gene present in transformants to be expressed before the cells are exposed to antibiotic. This strategy eliminates the potential clone bias that can be introduced through liquid outgrowth protocols. The grown colonies are picked for template preparation using the Qbot or QPix colony-picking robots (Genetix) and inoculated into 384-well blocks containing liquid media and incubated overnight with shaking. A small volume of bacterial growth from each clone is preserved in 25% glycerol and stored at -80∞C for genome finishing work and or clone distribution to the scientific community. Sequencing protocols are based on the di-deoxy sequencing method. To obtain paired sequence reads from opposite ends of each clone insert, two 384 well cycle sequencing reaction plates are prepared from each plate of plasmid template DNA. Sequencing reactions are carried out using Big Dye Terminator chemistry version 3.1 (Applied Biosystems) and standard M13 or custom forward and reverse primers. We have achieved excellent sequencing results using Big Dye terminator chemistry with reaction volumes and terminator concentrations substantially lower than those recommended by the manufacturer. Reaction mixtures, thermal cycling profiles, and electrophoresis conditions have been optimized to reduce the volume of the Big Dye Terminator mix to 1/16th of that recommended by the manufacturer and to extend read lengths on the AB 3730xl sequencers. We continue our efforts

in the reduction of the Big Dye Terminator mix and overall reaction volume that result in substantial cost savings. The protocols for 1/32 and 1/64 reactions are being developed by the R&D team. Sequencing reactions are set-up by the Biomek FX (Beckman) pipetting workstations. The robots are used to aliquot templates and to combine them with the reaction mixes consisting of deoxy- and fluorescently labeled dideoxynucleotides, the Taq thermostable DNA polymerase, sequencing primers, and reaction buffer. The template and reaction plates are bar coded and tracked by the bar code readers on the Biomek FX work-stations to assure error-free template and reaction mix transfer. The control software for this instrument will be directly integrated with our new LIM system. Thirty to forty consecutive cycles of linear amplification steps are performed on MJ Research Tetrads or 9700 thermal cyclers (Applied Biosystems). Reaction products are efficiently precipitated by isopropanol, dried at room temperature, and stored at 4∞C or resuspended in water and sealed. As sequencing machines become available, a sample sheet for each plate is automatically generated upon scanning the plate's barcode. The plates are then transferred to one of the AB3730xl DNA Analyzers for electrophoresis. The current polymers and software allow for 12 electrophoresis runs per day on an AB 3730xl with a set-up time of less than one hour.

Sequencing facilities and technology. In early 2003, TIGR and its affiliates established a new, large-scale sequencing facility, the J. Craig Venter Science Foundation Joint Technology Center (JTC). The JTC is approximately one-half mile from the current TIGR campus, and is managed by an oversight committee including representatives from TIGR and its two scientific affiliates, TCAG and IBEA. The heart of the facility is a laboratory with 100 ABI 3730xl DNA sequencers, which give it a capacity of 40,000,000 lanes per year. The high-throughput production team at the JTC is led by Tamara Feldblyum, who has been the Director of the TIGR Sequencing Facility since 1998. All sequence data for TIGR projects is delivered directly to TIGR's databases and file systems. The JTC DNA sequencing facility is one of the largest and the most complex in the world, and concurrently works on approximately 75 projects (see http://www.venterscienceجtc.org for a current list). More than 3.4 million sequencing reactions were performed at TIGR in 2001, and over 6,000,000 reactions were performed in 2002, representing a 76% increase in a single year. During the winter of 2002-3, this sequencing capacity was increased 250%, to 15,000,000 sequences/year through the acquisition of 24 new AB 3730xls. The overall average sequencing success rate for 2002 was 82.5 percent, and ranged between 81% and 90% depending on the G+C context of the DNA and the complexity of the genome. This number reflects high throughput sequence production from whole genome shotgun sequencing projects, BAC based projects, cDNA, and BAC ends sequencing as well as R&D and genome closure efforts. In 2003, 100 of the latest-generation AB 3730xl sequencers were installed at the new JTC facility, bringing the total sequencing capacity to 40,000,000 sequence reads per year. In tandem, the library construction, template production and sequencing reaction preparation processes were also upgraded to facilitate the 250% growth in sequencing capacity. Currently, the average usable read length for each sequence read is 803 bases, with a sequencing success rate of approximately 85%.

3. Bioinformatics Development at TIGR

TIGR's Bioinformatics Department, made up of over 70 scientists and engineers, has accumulated extensive experience in the course of annotating 26 published complete bacterial and eukaryotic genomes (e.g.,). As part of these projects, TIGR has developed software tools,

database structures, and computational resources designed to aid in the assembly, annotation, and analysis of genome sequence data. They have been described in publication and released for use by non-profit entities (for a list see http://www.tigr.org/softlab/). Development of new software and infrastructure is an ongoing process. TIGR has also developed tools for the release of genomic sequence information and annotation on the web. This includes tools for providing access to TIGR's annotation system for collaborators as well as access to sequence information and annotation to the outside (see http://www.tigr.org/tdb/). These databases include the Comprehensive Microbial Resource, TIGRfams, TIGR gene indices, and databases for multiple species or groups of species (e.g., plants, protists). A version of the Comprehensive Microbial Resource for single-celled eukaryotes is being developed. TIGR's prior and ongoing work on protist genomics (e.g., Entamoeba species, the Apicomplexans and the *Trypanosomes)* has led to the development of significant experience in analyzing protist genomes and gene functions which should assist in designing tools for analyzing the *A. castellanii* genome.

D. EXPERIMENTAL DESIGN AND METHODS

1. Whole genome shotgun (WGS) sequencing and assembly

Strain selection. *A. castellanii* Neff ATCC 30010 ("Acanthamoeba castellanii (Douglas) Page deposited as Acanthamoeba sp. Designation: Neff") will be used for this sequencing project. For the past two decades, this strain has been the subject of intensive biochemical and molecular biological research in the lab of one of the Co-PIs (Gray) (5). This work has largely focused on the structure and expression of the *A. castellanii* mitochondrial genome, which has been completely sequenced for this strain (6). An *A. castellanii* EST project has recently been launched, again using this particular strain, under the auspices of the "Protist EST Program" (PEP), an initiative partially funded by Genome Canada. For several amoebae, including *A. castellanii*, the Gray lab has noted significant strain-specific nucleotide sequence differences at the level of both nuclear and mitochondrial DNA. Such differences constitute a compelling practical consideration for sequencing the genome of the same strain of *A. castellanii* for which complete mitochondrial and extensive EST data are available.

DNA sequencing. We propose to sequence enough clones to produce an 8-fold sequence coverage of the genome. The amount of sequencing we propose to do is based on estimates of a genome size of 33 Mb. Approximately 50% of the sequencing will be achieved using a pHOS2 library with 6 – 8.0 kb inserts. The remaining sequencing will come from a combination of plasmids with larger average insert size and from linking libraries. Clones will be sequenced from both ends to produce pairs of linked sequences representing ~800 bp at the end of each insert. The randomness of the libraries will be individually assessed by comparing the redundancy in assemblies among each library individually and by comparing these to the theoretical redundancy based on the amount of sequence coverage from the library and the genome size. Once the randomness is determined for each library, the libraries for the bulk of the production sequencing will be selected to provide maximal clone coverage from random libraries.

Sequence Assembly. Assembly of the *Acanthamoeba* genome will be accomplished with the Celera Assembler, which is currently in use at TIGR. The Celera Assembler was used to assemble the data from the WGS projects for *Drosophila melanogaster* and humans. This capability is more than sufficient for the assembly of the *Acanthamoeba* genome, which is ~100 fold smaller than the human genome. However we will make use of other assembly programs if

they become available and are shown to be an improvement over the Celera Assembler (comparisons of assemblers are an ongoing area of research in TIGR's Bioinformatics Department). One possible such program is the new version of TIGR Assembler which is being developed at TIGR and which will include many of the features of the Celera Assembler. Assemblies will be run every month (using Celera Assembler until an alternative is found to be better) and the resulting contigs will be released to TIGR's web site.

2. **Genome annotation and analysis**

The annotation of the *Acanthamoeba* genome will use a pipeline that has been developed at TIGR for other eukaryotic genome projects (e.g., *E. histolytica*, *A. thaliana, P. falciparum*). This highly automated process consists of a set of eukaryotic annotation tools that rapidly identifies protein encoding genes and other features in genomic sequences. Contigs of any size will be analyzed as described below to find and characterize genes and other features of biological interest and the results will be stored in a relational database. The analyses will then be presented in a graphical user interface that will be available to the scientific community over the World Wide Web and submitted to Genbank/NCBI.

Identification of protein coding regions (CDS). For the initial identification of protein coding sequences (CDS), we will use the Glimmer algorithm which was developed at TIGR and provides a method for gene identification using interpolated Markov models. A eukaryotic version of Glimmer, called Glimmer M, was developed as a gene-finding tool for finding *Plasmodium* genes in genomic sequence, and is now also trained and available for *A. thaliana*, rice, and *T. parva*. Glimmer M will be trained using all the known *Acanthamoeba* gene and EST sequences and will be used as one the gene finder for this *Acanthamoeba* project. Additional gene finders that can be trained on EST data include Phat which will also be used. Such training should prove to be very efficient at producing efficient genefinding as there are significant numbers of Acanthamoeba EST sequences available (as of this writing, >12,000 ESTs representing >6000 unique clusters are available through PEPdb, the database of the Protist EST Program). All predicted CDS will be searched against protein sequence databases using a variety of tools (described below). In addition, regions of the genome without predicted coding regions and Glimmer predictions with no database match are re-evaluated using blastx as the initial search; new genes are then extrapolated from regions of alignment. Genomic *Acanthamoeba* sequences will be searched against EST and cDNA data using Sim4 and against protein databases using DPS/NAP. Because both of these algorithms take splicing into account, the resulting alignments give a much better representation of exon-intron boundaries than standard BLAST analyses. In addition to the homology-based analysis described above, several *ab initio* gene prediction algorithms will be used, including Genscan. A more accurate approach is to use sequence homology and *ab initio* predictions (from multiple gene finders) as input to a separate program that combines all the evidence and produces one (or more) gene predictions that represent a synthesis of all available inputs. This approach, called a Combiner, has been demonstrated to lead to better predictions than any single gene finder alone for plant genome data, and TIGR's studies on *P. falciparum* (unpublished data) indicate it also improves performance on protist genomes. In order to train the Combiner, TIGR curators will assemble a set of highly accurate known genes to use as training for the Combiner: this data, together with the pipelines' outputs and homology data, will be used by the Combiner to assign weights to all the various pieces of evidence. Genes coding for some untranslated RNAs will be identified by

database searches at the nucleotide level. A search for tRNA genes will be performed using tRNAScan-SE. Searches for methylation-guide (box C/D) snoRNA genes will use the computational screen developed by Lowe & Eddy (27), further trained on *A. castellanii* snoRNA sequences as they are revealed.

Functional prediction/annotation. TIGR's Bioinformatics Department has developed a variety of robust tools for sequence similarity searches and gene function prediction for putative CDS. These have been used to aid in the manual annotation of multiple prokaryotic and eukaryotic genomes and in the automated annotation of genomes that are not undergoing closure. Searches of the CDS are first performed with blastp. Gene identification is facilitated by searching against a database of non-redundant proteins (nraa) developed at TIGR and curated from public archives. Searches matching entries in nraa have the corresponding role, gene common name, percent identity and similarity of match, the pairwise sequence alignment, and taxonomy associated with the match assigned to the predicted coding region and stored in the database. In order to enhance our ability to make potential gene identifications, approaches and tools based on multiple sequence alignment and family building are employed. Paralogous gene families are created from multiple sequence alignments made with the target genome's predicted amino acid sequences. The multiple sequence alignments are used to group similar proteins into families for verification of annotation and identification of family members perhaps not recognized by simple pairwise alignment. The protein models generated by the searches and predictions will be further searched against Markov model (HMM) databases including PFAM and TIGRFAMs.

TIGR's Bioinformatics Department is continually looking for additional tools to add to the annotation process. Examples include using phylogenetic analysis to aid in functional predictions, implementing so-called non-homology methods for functional annotation such as phylogenetic profiling and Rosetta stone methods. These methods have begun to be used in the manual annotation of bacterial genomes. Efforts are underway to automate these processes so that they can be added directly to the annotation pipelines. Finally, each putatively identified gene is assigned to one of 102 role categories adapted from Riley. Work is under way to incorporate the Gene Ontology (GO) system into the annotation pipeline; much of the GO system is already in place for eukaryotic annotation.

Repeat analysis. Repeated sequences in the genome will be identified using an efficient algorithm based on suffix trees. The initial set of repeats is further processed to group repeats into classes, which are used to guide both assembly and annotation (46). Large repeats will be compared using the MUMmer software, which allows for alignments of very large segments of DNA, and has been used to identify and characterize duplications (whole genome and segmented) in a variety of specie.

Gene indices. The *Acanthamoeba* EST sequences will be assembled in the TIGR gene index database to provide a representation of mRNA sequences assembled from overlapping EST sequences. This resource will be valuable during the annotation review stage of the project, serving a resource for scientists to use in validating electronic gene calls and to verify the exon/intron boundaries called by Glimmer M.

Comparative and evolutionary studies. We also plan to conduct a series of comparative and evolutionary analyses of the sequence data for *A. castellanii.* We have developed a series of methods for phylogenomic analyses that allow the characterization of evolutionary events in the history of a genome as well as improve the understanding of the genome itself through evolutionary analysis. These include methods to classify multigene families into subfamilies (11, 12), identify recent duplications in the history of a genome, detect possible lateral gene transfer events such as those from organellar genomes (19) (29), to distinguish possible gene-transfer events from gene loss and evolutionary rate variation (38), to characterize genome duplications and rearrangements (10), and to study species evolution (24). In addition, we have begun to use methods that allow the grouping of genes by features other than sequence similarity (the so-called non-homology methods). For example, we have used phylogenetic profiling, a method for grouping genes based on their distribution patterns across species (35) to study the genes involved in energy metabolism in *Chlorobium tepidum* and pathogenicity in *S. pneumoniae*. Among the questions we will address in *A. castellanii* using these and other methods are:

1) Which genes are orthologous to proteins from amoebae and other eukaryotes?
2) What genes are present in *Ac* and *Dd* but absent in *Eh* ?
Similarly, which genes are present in Ac and Eh but not Dd? And how do these relate to gene complements in the amoebic parasites?
3) Are there any gene families that have expanded in the *Acanthamoeba* lineage?
4) Can we identify genes of plastid or mitochondrial descent in the nuclear genome?
5) Are there any large-scale intragenomic duplications?
6) What is the phylogenetic position of *A. castellanii ?* Previous studies have been based on only a limited set of genes. We will compare and contrast the phylogenetic trees of a set of genes found in all eukaryotic species for which complete genomes are available.
7) What are the DNA repeat elements like in the genome (as identified by sequence comparisons)? Are there any new classes of repeats that may be regulatory elements or new transposons?
8) Is there any evidence of unique expression mechanisms (e.g., examples of DNA editing, unusual splicing events, specific categories of small RNAs) or of unusual genome arrangements?

3. **Data release and clone availability**

We plan to make all of the data from this project available as soon as is possible to the research community. Sequence data and assembly data will be released regularly to TIGR's web and FTP sites. Since genome assembly is planned to be conducted approximately bi-monthly, assembly data will be released approximately bi-monthly. If additional assembling is carried out, the assemblies will be released more frequently. Sequence data will be released weekly. All reasonable efforts will be made to make clones available to the research community upon request. However, the large number of clones that are to be generated means that the release of clones may be too costly in some cases.

TIGR has also developed graphical web pages for viewing WGS annotation, which are now being used as part of the analysis of multiple eukaryotic genomes. This display is constantly being improved and enhanced, and illustrates one of the means by which we will make the *A. castellanii* annotation available. The TIGR *A. castellanii* annotation that is presented to the web will initially be in this format and will include information on assemblies, gene models, database

matches (ESTs, cDNAs and proteins) and putative role category assignment. The TIGR website will also include multiple downloading options, including the ability to download DNA sequence data, predicted protein data, as well as functional annotation information. In addition, we plan to release the *A. castellanii* genomic annotation as part of the Eukaryotic Genome Resource (ER), which is being developed currently at TIGR. The ER is modeled on the Comprehensive Microbial Resource (CMR), a web based resource created and maintained by TIGR for presentation and analysis of prokaryotic genome data (http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl). The CMR automatically updates and re-analyzes information relating to gene function, biological role, placement into gene families, three-dimensional structure, and links to other databases. In addition, it allows the user to make meaningful inter- and intra-genomic comparisons.

4**. Outreach**

Given consideration of the relatively large size of the *A. castellanii* genome and the relatively high costs of both finishing, closure and manual annotation in an environment of limited funding we have endeavored to keep the overall project costs as low as possible. Given the disseminated nature of the *A. castellanii* community and its focuses on quite disparate aspects of *A. castellanii* biology (see attached letters of support) we feel that providing ready access to a high quality genome assembly is the best form of outreach that we can provide without proposing inordinately expensive measures to perform a community based extensive manual annotation of a relatively large genome. Throughout the lifetime of the project however we will have constant contact with various interested parties from the *Ac*, microbiological and protist communities, we will generate a steering committee one of whose tasks it will be to interface with the Ac community. This notion of a steering committee has worked very well within the *E. histolytica* genome project for providing a forum for members of the community to contribute to the genome project either in terms of suggestions, annotations or items of interest for a publication describing the genome. The genome is being annotated with assistance from almost the entire community however the community interactions and contributions have mainly been via email and file transfer protocols. Members of the steering committee would represent leaders from their communities and likely attend various conferences with interests in the *Ac* genome and present progress reports and updates as well as facilitating contributions from other sources. We are open to the idea of seeking funding for a genome users workshop in conjunction with the Protist EST Program (PEP) consortium where members of the community could review and add to aspects of the final annotation and discuss the impacts of the genome project on their work. Given the reasons cited above however we are reluctant to request funding for such an undertaking at this stage.

5.**The overall project timeline and Integration**.

The entire project timelines will be carried out in consultation with the Co PI

Months 1-3. Preparation and test sequencing of additional sized plasmid libraries for Ac. Generation of an Ac gene index as a complementary resource to the genome project for the broader community. Generation of a steering committee from within members of the Ac community as well as evolutionary protistologists to assist in community outreach and also to assist with the generation of a manuscript describing the findings of the genome sequence.

Months 3-9 Sequencing and assembly of the various plasmid based libraries to 8-fold coverage of the *Ac* genome. Continuous quality control of the libraries, the presence of contaminating sequences and correct assembly of the genome. Monthly data release of the assembled data through a specific designated web site. Training of a variety of gene finding algorithms using the Ac gene index and published Ac gene sequences from Genbank.

Months 9-12 Generation of final genome assembly and an initial automated annotation, Interaction and consultation with the Ac and broader community to assist in improvement of the automated annotation by manually correcting annotation errors identified by members of the Ac community. Generation and release of the final improved annotation onto a dedicated web site. Working with the steering committee to identify items of biological interest from the genome sequence data and identification of members of the Ac community to assist in preparation of the final manuscript outlining the findings of the genome sequence.

REFERENCES

1. **Abd, H., T. Johansson, I. Golovliov, G. Sandstrom, and M. Forsman.** 2003. Survival and growth of Francisella tularensis in Acanthamoeba castellanii. Appl Environ Microbiol **69:**600-6.
2. **Allen, P. G., and E. A. Dawidowicz.** 1990. Phagocytosis in Acanthamoeba: I. A mannose receptor is responsible for the binding and phagocytosis of yeast. J Cell Physiol **145:**508-13.
3. **Bisson, R., S. Vettore, E. Aratri, and D. Sandona.** 1997. Subunit change in cytochrome c oxidase: identification of the oxygen switch in Dictyostelium. Embo J **16:**739-49.
4. **Brown, M. R., A. W. Smith, J. Barker, T. J. Humphrey, and B. Dixon.** 2002. E. coli O157 persistence in the environment. Microbiology **148:**1-2.
5. **Bullerwell, C. E., M. N. Schnare, and M. W. Gray.** 2003. Discovery and characterization of Acanthamoeba castellanii mitochondrial 5S rRNA. Rna **9:**287-92.
6. **Burger, G., I. Plante, K. M. Lonergan, and M. W. Gray.** 1995. The mitochondrial DNA of the amoeboid protozoon, Acanthamoeba castellanii: complete sequence, gene content and genome organization. J Mol Biol **245:**522-37.
7. **Byers, T. J.** 1986. Molecular biology of DNA in Acanthamoeba, Amoeba, Entamoeba, and Naegleria. Int Rev Cytol **99:**311-41.
8. **Cengiz, A. M., N. Harmis, and F. Stapleton.** 2000. Co-incubation of Acanthamoeba castellanii with strains of Pseudomonas aeruginosa alters the survival of amoeba. Clin Experiment Ophthalmol **28:**191-3.

9. **Clarholm, M.** 2002. Bacteria and protozoa as integral components of the forest ecosystem--their role in creating a naturally varied soil fertility. Antonie Van Leeuwenhoek **81:**309-18.
10. **Eisen, J. A.** 2000. Assessing evolutionary relationships among microbes from whole-genome analysis. Curr Opin Microbiol **3:**475-80.
11. **Eisen, J. A.** 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res **8:**163-7.
12. **Eisen, J. A., K. S. Sweder, and P. C. Hanawalt.** 1995. Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions. Nucleic Acids Res **23:**2715-23.
13. **Essig, A., M. Heinemann, U. Simnacher, and R. Marre.** 1997. Infection of Acanthamoeba castellanii by Chlamydia pneumoniae. Appl Environ Microbiol **63:**1396-9.
14. **Ettinger, M. R., S. R. Webb, S. A. Harris, S. P. Mclninch, C. G. G, and B. L. Brown.** 2003. Distribution of free-living amoebae in James River, Virginia, USA. Parasitol Res **89:**6-15.
15. **Finlay, B. J., and T. Fenchel.** 1999. Divergent perspectives on protist species richness. Protist **150:**229-33.
16. **Fritsche, T. R., M. Horn, S. Seyedirashti, R. K. Gautom, K. H. Schleifer, and M. Wagner.** 1999. In situ detection of novel bacterial endosymbionts of Acanthamoeba spp. phylogenetically related to members of the order Rickettsiales. Appl Environ Microbiol **65:**206-12.
17. **Gaze, W. H., N. Burroughs, M. P. Gallagher, and E. M. Wellington.** 2003. Interactions between Salmonella typhimurium and Acanthamoeba polyphaga, and Observation of a New Mode of Intracellular Growth within Contractile Vacuoles. Microb Ecol.
18. **Greub, G., B. La Scola, and D. Raoult.** 2003. Parachlamydia acanthamoeba is endosymbiotic or lytic for Acanthamoeba polyphaga depending on the incubation temperature. Ann N Y Acad Sci **990:**628-34.
19. **Hoffmeister, M., and W. Martin.** 2003. Interspecific evolution: microbial symbiosis, endosymbiosis and gene transfer. Environ Microbiol **5:**641-9.
20. **Horowitz, J. A., and J. A. Hammer, 3rd.** 1990. A new Acanthamoeba myosin heavy chain. Cloning of the gene and immunological identification of the polypeptide. J Biol Chem **265:**20646-52.
21. **Inglis, T. J., P. Rigby, T. A. Robertson, N. S. Dutton, M. Henderson, and B. J. Chang.** 2000. Interaction between Burkholderia pseudomallei and Acanthamoeba species results in coiling phagocytosis, endamebic bacterial survival, and escape. Infect Immun **68:**1681-6.
22. **Jeon, T. J., and K. W. Jeon.** 2003. Characterization of sams genes of Amoeba proteus and the endosymbiotic X-bacteria. J Eukaryot Microbiol **50:**61-9.
23. **King, C. H., E. B. Shotts, Jr., R. E. Wooley, and K. G. Porter.** 1988. Survival of coliforms and bacterial pathogens within protozoa during chlorination. Appl Environ Microbiol **54:**3023-33.
24. **Kunin, V., and C. A. Ouzounis.** 2003. The balance of driving forces during genome evolution in prokaryotes. Genome Res **13:**1589-94.
25. **Kurland, C. G., B. Canback, and O. G. Berg.** 2003. Horizontal gene transfer: a critical view. Proc Natl Acad Sci U S A **100:**9658-62.
26. **La Scola, B., and D. Raoult.** 2001. Survival of Coxiella burnetii within free-living amoeba Acanthamoeba castellanii. Clin Microbiol Infect **7:**75-9.
27. **Lowe, T. M., and S. R. Eddy.** 1999. A computational screen for methylation guide snoRNAs in yeast. Science **283:**1168-71.
28. **Maciver, S. K., and P. J. Hussey.** 2002. The ADF/cofilin family: actin-remodeling proteins. Genome Biol **3:**reviews3007.
29. **Martin, A. C., and D. G. Drubin.** 2003. Impact of genome-wide functional analyses on cell biology research. Curr Opin Cell Biol **15:**6-13.
30. **Martin, W., and M. Muller.** 1998. The hydrogen hypothesis for the first eukaryote. Nature **392:**37-41.
31. **Michel, R., K. D. Muller, and R. Hoffmann.** 2001. Enlarged Chlamydia-like organisms as spontaneous infection of Acanthamoeba castellanii. Parasitol Res **87:**248-51.

32. **Miltner, E. C., and L. E. Bermudez.** 2000. Mycobacterium avium grown in Acanthamoeba castellanii is protected from the effects of antimicrobials. Antimicrob Agents Chemother **44:**1990-4.
33. **Newsome, A. L., T. M. Scott, R. F. Benson, and B. S. Fields.** 1998. Isolation of an amoeba naturally harboring a distinctive Legionella species. Appl Environ Microbiol **64:**1688-93.
34. **Nixon, J. E., A. Wang, J. Field, H. G. Morrison, A. G. McArthur, M. L. Sogin, B. J. Loftus, and J. Samuelson.** 2002. Evidence for lateral transfer of genes encoding ferredoxins, nitroreductases, NADH oxidase, and alcohol dehydrogenase 3 from anaerobic prokaryotes to Giardia lamblia and Entamoeba histolytica. Eukaryot Cell **1:**181-90.
35. **Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates.** 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A **96:**4285-8.
36. **Rodriguez-Zaragoza, S.** 1994. Ecology of free-living amoebae. Crit Rev Microbiol **20:**225-41.
37. **Rodriguez-Zaragoza, S., and A. Magana-Becerra.** 1997. Prevalence of pathogenic Acanthamoeba (Protozoa:Amoebidae) in the atmosphere of the city of San Luis Potosi, Mexico. Toxicol Ind Health **13:**519-26.
38. **Salzberg, S. L., O. White, J. Peterson, and J. A. Eisen.** 2001. Microbial genes in the human genome: lateral transfer or gene loss? Science **292:**1903-6.
39. **Schuster, F. L.** 2002. Cultivation of pathogenic and opportunistic free-living amebas. Clin Microbiol Rev **15:**342-54.
40. **Segal, G., and H. A. Shuman.** 1999. Legionella pneumophila utilizes the same genes to multiply within Acanthamoeba castellanii and human macrophages. Infect Immun **67:**2117-24.
41. **Shaw, J. J., F. Dane, D. Geiger, and J. W. Kloepper.** 1992. Use of bioluminescence for detection of genetically engineered microorganisms released into the environment. Appl Environ Microbiol **58:**267-73.
42. **Simek, K., J. Nedoma, J. Pernthaler, T. Posch, and J. R. Dolan.** 2002. Altering the balance between bacterial production and protistan bacterivory triggers shifts in freshwater bacterial community composition. Antonie Van Leeuwenhoek **81:**453-63.
43. **Stechmann, A., and T. Cavalier-Smith.** 2003. The root of the eukaryote tree pinpointed. Curr Biol **13:**R665-6.
44. **Steenbergen, J. N., H. A. Shuman, and A. Casadevall.** 2001. Cryptococcus neoformans interactions with amoebae suggest an explanation for its virulence and intracellular pathogenic strategy in macrophages. Proc Natl Acad Sci U S A **98:**15245-50.
45. **Tharavathi, N. C., and B. B. Hosetti.** 2003. Biodiversity of algae and protozoa in a natural waste stabilization pond: a field study. J Environ Biol **24:**193-9.
46. **Volfovsky, N., B. J. Haas, and S. L. Salzberg.** 2001. A clustering method for repeat analysis in DNA sequences. Genome Biol **2:**RESEARCH0027.
47. **Wang, X., and D. G. Ahearn.** 1997. Effect of bacteria on survival and growth of Acanthamoeba castellanii. Curr Microbiol **34:**212-5.
48. **Weekers, J. F.** 1993. [A new kind of keratoprosthesis]. Bull Soc Belge Ophtalmol **247:**25-7.
49. **Weekers, P. H., A. M. Engelberts, and G. D. Vogels.** 1995. Bacteriolytic activities of the free-living soil amoebae, Acanthamoeba castellanii, Acanthamoeba polyphaga and Hartmannella vermiformis. Antonie Van Leeuwenhoek **68:**237-43.
50. **Wieltschnig, C., U. R. Fischer, A. K. Kirschner, and B. Velimirov.** 2003. Benthic bacterial production and protozoan predation in a silty freshwater environment. Microb Ecol **46:**62-72.
51. **Winiecka-Krusnell, J., K. Wreiber, A. von Euler, L. Engstrand, and E. Linder.** 2002. Free-living amoebae promote growth and survival of Helicobacter pylori. Scand J Infect Dis **34:**253-6.
52. **Zusman, T., G. Yerushalmi, and G. Segal.** 2003. Functional similarities between the icm/dot pathogenesis systems of Coxiella burnetii and Legionella pneumophila. Infect Immun **71:**3714-23.