COMPARATIVE GENOME EVOLUTION

MODIFIED PROPOSAL

Introduction

Analysis of comparative genomic sequence information from a well-chosen set of organisms is, at present, one of the most effective approaches available to advance biomedical research. The following document describes rationales and plans for selecting targets for genome sequencing that will provide insight into a number of major biological questions that broadly underlie major areas of research funded by the National Institutes of Health, including studies of gene regulation, understanding animal development, and understanding gene and protein function.

Genomic sequence data is a fundamental information resource that is required to address important questions about biology: What is the genomic basis for the advent of major morphogenetic or physiological innovations during evolution? How have genomes changed with the addition of new features observed in the eukaryotic lineage, for example the development of an adaptive immune system, an organized nervous system, bilateral symmetry, or multicellularity? What are the genomic correlates or bases for major evolutionary phenomena such as evolutionary rates; speciation; genome reorganization, and origins of variation? Another vital use to which genomic sequence data are applied is the development of more robust information about important non-human model systems used in biomedical research, i.e. how can we identify conserved functional regions in the existing genome sequences of important non-mammalian model systems, so that we can better understand fundamental aspects of, for example, gene regulation, replication, or interactions between genes?

NHGRI established a working group to provide the Institute with well-considered scientific thought about the genomic sequences that would most effectively address these questions. The proposal described below is a modification of the original justifications and discussions of the Comparative Genome Evolution Working Group, based on the outcomes of discussions held by a committee charged with coordinating that working group's report and a working group on annotating the human genome, as well as discussions of the National Advisory Council for Human Genome Research.

The proposal is organized into three sections: (1) Metazoan origins of the human genome; (2) Linking genomic change to life history and behavior; and (3) Protist origins of the human genome. Each section consists of a detailed discussion

followed by proposals for specific examples of organisms that would address the issues raised in the discussion. Candidates for sequencing targets were chosen by NHGRI from among those examples. The working group's proposal mentions many specific organisms in various contexts, but only a subset was selected for sequencing by NHGRI in this round of consideration. The selected sequencing targets are:

The lamprey *Petromyzon marinus*
The nematode worm *Pristionchus pacificus*
The nematode worm *Trichinella spiralis*
The snail *Biomphalaria glabrata*
Hydra magnipapillata
Trichoplax adhearens
Oxytricha trifallax (macronucleus plus equivalent micronucleus)
Monosiga ovata
Physarum polycephalum

It is important to emphasize that as new sequence data accumulate, new scientific conclusions will emerge, new opportunities will present themselves, and our understanding of the overarching rationale described above will mature. One of the objectives of the NHGRI's current approach to choosing new sequencing targets is to remain flexible enough to adjust to new facts and opportunities. In addition, as the overarching goal is to address broad and fundamental questions, the implementation effort at a specific point in time, will only partially address it. Therefore, it is anticipated that the specific plans and justifications below will develop and change over time.

Comparative Genome Evolution Working Group: Three proposals for genome sequencing

Table of Contents

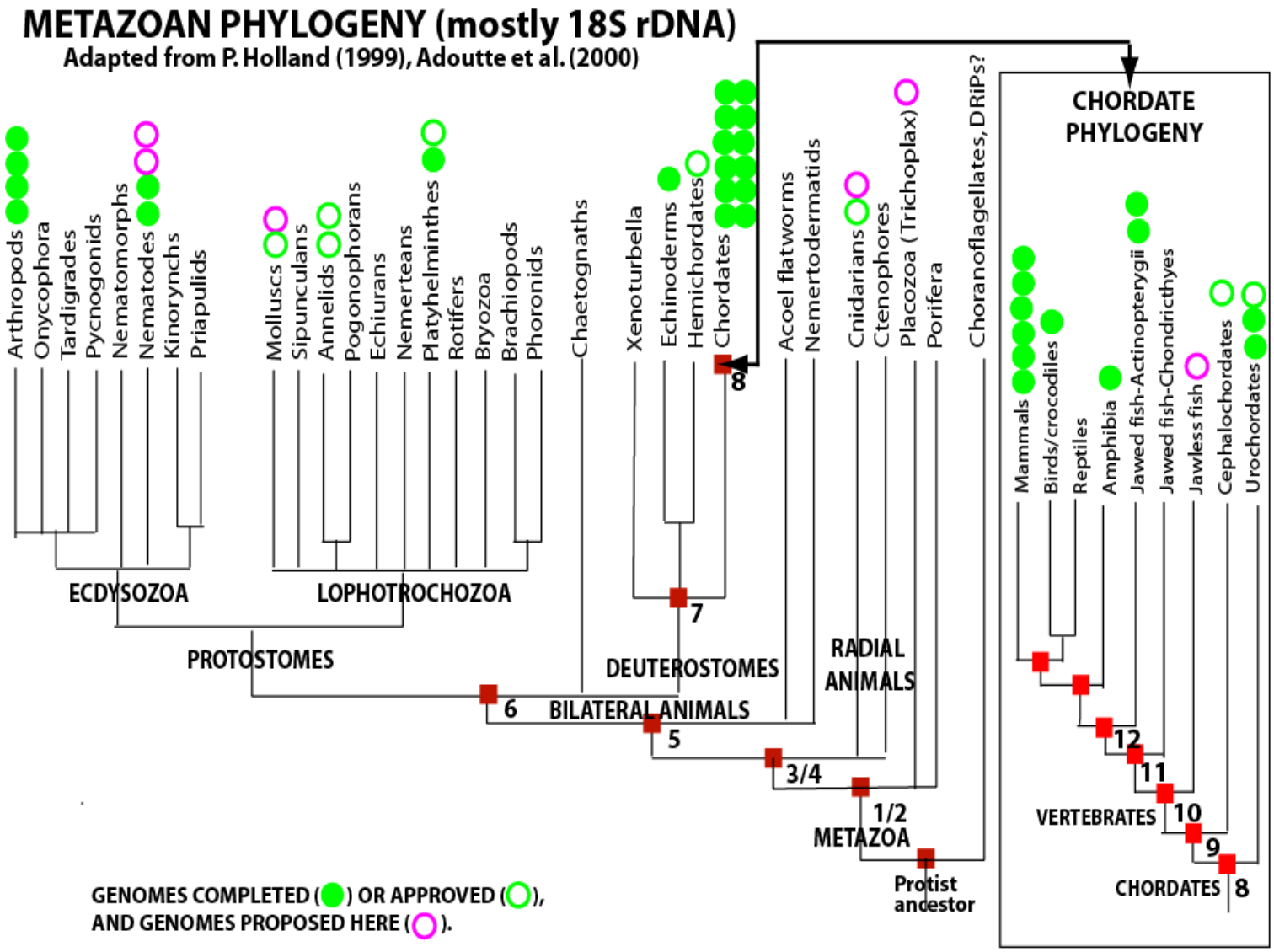
Section 1: Metazoan origins of the human genome

A. The question

How, when, and why in the course of evolution, did the human genome acquire its current composition and organization? In this context, 'composition' is used to mean the genes, cis-regulatory sequences, and replication origins encoded within the human genome, and 'organization' is used to mean the gene clusters, order and chromosomal position of genes, and intron-exon arrangements of the genome. This will help us understand how the genome functions to produce the most critical and typical features of complex multicellular life, including human biology.

It is well known that many sequences of the human genome originated long before humans themselves. For example, some human sequences strongly resemble sequences found in prokaryotes, reflecting a conservation of sequence during the 2-3 billion years since a last common ancestor. Other human sequences strongly resemble those of eukaryotic protists, reflecting a conservation of sequence of perhaps 1-2 billion years from a protist ancestor. Yet other human sequences strongly resemble those only of other animals, reflecting conservation over hundreds of millions of years of metazoan evolution. The conservation of protein-coding and some RNA-coding sequences over long periods is a fact made undeniable by the results of sequencing efforts to date, and the attribution of encoded functions to those human genes has routinely relied on what is known about the function of the similar genes in other organisms, namely, on conservation.

Figure 1. Metazoan phylogeny



The most effective approach available today to obtaining answers about deep sequence conservation in the human genome, in a systematic manner, is comparative genomics — the comparison of human sequences to sequences of genomes of organisms in taxa that diverged at different times from the human lineage. Each divergence point is a “node.” Examination of the metazoan tree (see Figure 1, with nodes numbered) shows that it contains only about 12 nodes on the human lineage below the grade of ‘fish’, back to the sponges and placozoans at the very base of the tree. That final node would be the ancestor of all Metazoa, dating back perhaps 0.6-1.2 Bya (estimates vary widely). Beyond that time, the protist ancestry of the human genome would be reached (e.g., choanoflagellates, DRIPs, etc), as will be discussed in part of the proposal Section 3. New sequencing projects of a representative from each of the 12 nodes is not warranted, however, because some of the have already been completed or are underway (e.g., *Ciona*, amphioxus, sea urchin, some protostomes). The working group’s proposal calls for sequencing to fill in the missing nodes (such as nodes 1-6, 10, 11), and to acquire additional data in some major nodes that are already partially studied. Furthermore, the modified proposal does not address nodes within the tetrapod vertebrates (including amphibia), since these groups are already well sampled for genome sequencing (see the other major section on Annotating the Human Genome) and that information is readily available for use in comparative genomics.

Each of the 12 nodes represents a unique suite of major changes in biology -- animal development, morphology, physiology, endocrinology, and behavior. Each node is expectedly accompanied by major sequence changes in the genome, namely, changes of coding regions, regulatory regions, gene family expansions and diversifications, and gene organization. Some but not all changes will, we expect, have become conserved at some time in the past, thereafter carried forward in the human lineage to the present. This comparative work will illuminate, not just times of origin, but also gene function, gene identification, and regulatory evolution (especially that mediated through gene clustering), and will assist studies of the evolution of development and post-genomic analyses of physiology.

B. The nodes and proposed examples of species to sequence

- **Node 12**: **Ray-finned fish as the sister group to tetrapods** (plus lungfish and coelocanth) [Divergence approximately 420 Mya]: Ray-finned fish share with tetrapods, such as human, all the true vertebrate characters, but also the domination of bone in the axial skeleton and head. The notable difference is that this lineage diverged before the vertebrate conquest of land, and several key characters are missing, notably a true lung, walking legs, extensive limb musculature, axially-connected pelvic apparatus, middle ear ossicles for airborne sound detection and (presumably) adaptations for defense against airborne pathogens.

Comparison of tetrapod and ray-finned genomes will help elucidate genomic correlates of these adaptations. The >20,000 species of ray-finned fish consist primarily of teleosts, which are already well represented for comparative genomics by pufferfish (*Fugu*, *Tetraodon*) and zebrafish (*Danio*), and the stickleback (which has been approved for sequencing). However, a complication for these teleost species is that they share the remnants of an ancient genome duplication dating to early teleost evolution. This genome duplication has been followed by loss of different genes in each species (e.g. loss of different NK homeobox genes and melanocortin receptors in *Danio* and *Fugu*) causing problems for comparison. An example of a species that would address these factors would be of a basal (holostean) ray-finned fish that diverged prior to genome duplication, such as *Polypterus senegalus*. These species are not put forward at this time, but deserve future consideration.

- **Node 11: Common ancestor to all jawed vertebrates** [Divergence approximately 450 Mya]: The earliest evolutionary divergence within the jawed vertebrates is between cartilagenous fish (sharks, rays) and bony 'fish' (including human). Immediately prior to this node we see the evolution of true vertical biting jaws, strong cephalization, bilateral sense organs such as paired olfactory capsules, a well developed cerebellum, actively sustained motility, two sets of paired fins developing in a way similar to tetrapod fore- and hind-limbs, expanded steroid hormone usage, a complex adaptive immune system, myelinated nerves, four Hox clusters, a full array of Wnts, TGFbeta ligands, IGFs, etc. It is clear that the major gene family expansions that occurred in early vertebrate evolution were completed at this node. The sharks and rays---for example, *Scyliorhinus canicula*--- are thus important to consider as targets for sequencing. However, the very large size of many of the genomes in this group of animals tempers enthusiasm for choosing one at present.

- **Node 10**: **Jawless fish as basal vertebrates** [Divergence in the Cambrian >520 Mya?): This node represents the earliest point at which a true vertebrate can be identified, and represents a big step in anatomical and physiological complexity from the cephalochordate to the vertebrate grade. Principal innovations include neural crest derivatives such as a peripheral and enteric nervous system, cranial ganglia, and spinal ganglia, the sclerotome forming cartilaginous vertebrae, and the midbrain and forebrain. There are already tentative indications that genome complexity differs profoundly between these basal vertebrates and other vertebrates, as clear orthologues of all four human Hox clusters cannot be identified, and as other patterns of gene duplication differ. Although there is general agreement that lampreys belong to the sister group to jawed vertebrates, it is uncertain whether the hagfish (the only other jawless fish) should be included in this group or associated with a separate node. Thus, in terms of priority of genome sequencing, we strongly propose lamprey as the

appropriate basal vertebrate. Additionally, these can be studied developmentally and are a commercial concern in North America. The candidate species is the sea lamprey, *Petromyzon marinus*.

- **Node 9**: **Cephalochordates as non-vertebrate chordates** [Divergence time in the Cambrian, >545 Mya?]: The cephalochordates (amphioxus), as the closest living invertebrate relatives of vertebrates, share with ourselves the full suite of chordate characters: notochord, dorsal hollow nerve cord, gill slits, endostyle and post-anal tail. The nerve cord has dorsoventral specialization, although no cranial or trunk neural crest is present, while mesodermal somites form definitive myotomes and visceral muscle, but no sclerotome. Possession of basic chordate anatomy, but lack of the extensive specialization and elaboration typical of vertebrates, make this node of great significance for comparative genomics. It is already known that numerous vertebrate gene families are present as single copies in amphioxus (e.g. Hox cluster, many other homeobox genes, dystrophin, FGFR, IGF, VEGFR, COUP-TF), while gene linkages and clusters are often in an ancestral, underived, form in amphioxus (Hox, ParaHox, MHC-paralogy region). Of the 20 – 30 species, three are widely used in developmental and physiological research. One species, *Branchiostoma floridae*, is scheduled for genome sequencing by JGI. The working group agreed that this species is of high priority, particularly for potential insights into the conservation or diversification of physical gene linkages. As that sequencing advances, NHGRI should consider a second species (*Branchiostoma lanceolatum)*, but not as part of the present plan.

- **Node 8**: **Urochordates as basal chordates** [Divergence in the Cambrian, >520Mya?]: Urochordates (ascidians, larvaceans etc.) possess the main chordate traits of notochord, dorsal hollow nerve cord, gill slits, endostyle and post-anal tail. Like cephalochordates (node 9), the nerve cord has dorsoventral specialization, but the dramatic elaborations of vertebrate development, anatomy and physiology are not present. Paradoxically, the brain homologue seems more vertebrate-like than in cephalochordates, with a clear midbrain-hindbrain region expressing FGF, Pax2/5/8, and En genes. This recent finding stresses further why comparative genomics must embrace diversity to enable accurate reconstruction of human gene functions. As with cephalochordates, single genes usually correspond to vertebrate gene families, although current data suggest a greater degree of physical genomic rearrangement in urochordates than seen in cephalochordates. Currently, two species have had their complete genomes sequenced (*Ciona intestinalis, Ciona savignyi*), while a third (*Oikopluera dioica)* is scheduled for sequencing at the , JGI.

- **Node 7**: **The deuterostome ancestor** [Divergence in the Pre-Cambrian, >545 Mya?]: Chordates are in the supertaxon of Deuterosomes, along with the hemichordates and echinoderms. All other bilateral animals are

"protostomes". Thus, chordates have but two closely related phyla. Echinoderms are quite derived with a pentameric body organization, which complicates comparisons, but the bilateral hemichordates show partial chordate traits such as gill slits, a tail-like post-anal body extension, an endostyle (with thyroid ancestry?), and a proposcis pore (with pituitary ancestry?). The nervous system is diffuse with axon tracts, or is only locally condensed, in hemichordates and echinoderms, rather than being centralized. These non-chordate deuterostomes will be important for defining what was present in the immediate chordate ancestor. Both taxa are amenable to developmental and physiological research, with the echinoderms including some intensively studied model species. It is clear that the ancestor possessed a large suite of hox genes, and most (perhaps all) signaling pathways. The sequence of one echinoderm genome is almost complete (*Strongylocentrotus purpuratus*). Sample sequence is available for one other echinoderm (*Lytechinus variegates*), while one hemichordate (*Saccoglossus kowalevskii*) is already on the NHGRI's prioritized list of sequencing targets, although its sequencing is not yet underway.

- **Node 6**: **The ancestor of bilateral animals.** [Divergence in the Pre-Cambrian, >545Mya?]: The (Eu)Bilateria diverged into the deuterostomes (our lineage), and two extremely diverse protostomes groups, the Lophotrochozoa and Ecdysozoa (see Figure 1; other tree topologies are possible, but would not change the examples of the species that would be chosen to address the issues raised). At this node in our evolution, a full bilateral body plan was present with dorsoventral as well as anteroposterior organization, as was a through gut with mouth and anus. Cephalization with light receptors (under *pax-6* developmental control) may have been present, as well as many new mesodermal cell types. This ancestor had an expanded hox cluster, and most cell-cell signaling pathways, including characteristic protein-protein interactions (e.g. sog/dpp/twg/tolloid). There are, however, many uncertainties concerning the developmental characters of the bilaterial ancestor, because of insufficient data on genotype-phenotype links in diverse species. For example, the nervous system may have been diffuse, partially centralized or centralized and a pulsatile blood system may or may not have been present.

  Considering the extreme diversity of anatomy, embryology and physiology throughout the protostomes (ca. 25 phyla), the working group recommended that a deeper sampling of species is essential if the nature of this ancestor and its genome are to be understood. This will also be essential to distinguish conservation from convergent evolution when comparing gene functions and gene arrangements between vertebrates and protostomes. This is not an arcane problem, but one that is central in

biological sciences today, primarily because the protostomes include the common model systems *Drosophila melanogaster* and *Caenorhabditis elegans, and* it is very important to know when it is safe to extrapolate from these model systems to humans. In practical terms, the goals are to gain insight into the pathway of genomic evolution within arthropods (a major ecdysozoan phylum, including *Drosophila*), within nematodes (a major ecdysozoan phylum, including *Caenorhabditis*), at the base of the ecdysozoans, and within the largely unsampled lophotrochozoans.

- **Node 6.1: Arthropod genome diversity**:
  At present, full genome sequence is available, or will soon be available, for several *Drosophila* species. Outside of this dipteran insect species, sequencing has been completed or started for the genomes of several other insects, including mosquito (*Anopheles*), honeybee (*Apis*), and the flour beetle (*Tribolium*). Outside the insects, the only other arthropod for which whole genome sequencing has begun is the crustacean, *Daphnia*. For a group that is as species-rich and diverse as the arthropods (more than any other phylum), this set of species represents an extremely limited and biased sample. Virtually all of these organisms fall within one very small portion of the diverse phylum. In particular, there are at present no representatives from the phylogenetically primitive insects (hemimetabolous groups such as aphids, locusts, and silverfish). Furthermore, the three other main arthropod branches, namely, the chelicerates (spiders, scorpions, horseshoe crabs), myriapods (millipedes and centipedes), and crustaceans (crabs, copepods, brine shrimp, etc.), are not represented, except for *Daphnia*. Although *Daphnia* is an excellent starting point within the crustaceans because of its small genome and history of population genetic studies, it will be important to eventually include other crustaceans, given the level of diversity of this group. Sampling throughout the arthropods is essential if we are to properly predict the genome properties of the ancestral arthropod and then, by extension, the ancestral bilaterian.

  While there are ample reasons to sample a diversity of arthropods, the large genome size of so many members of this phylum represents a significant hurdle (see further discussion of this issue in Section 2). It will be useful to gather data to select various arthropods with a reasonable genome size, easy accessibility, and an already existing experimental community, especially to select candidate chelicerates and crustaceans at later time. For now, however, NHGRI should sequence a hemimetabolous insect that is a suitable “out-group” to all other insects (all “holometabolous”) currently being sequenced. An example species would be the Pea Aphid, *Acyrthosiphon pisum,* which also would be important from

the perspective of ecological, developmental, and agricultural investigations. Holometabouous insects also include important human disease vectors, a factor which should also be considered in selecting sequencing targets. To broaden the sampling within the hemimetabolous insects, there are a variety of other candidates that should be considered, a number of which have additional strengths as model systems with well-developed research communities, and which would also have the virtue of helping to annotate an existing genome: potential examples would be the wasp *Nasonia vitripennis* or *Sciara coprophila*, but there are many other possibilities that should also be considered.

It will also be important to examine an outgroup to the arthropods as a whole and to gather data on Tardigrades (water bears) in order to inform a recommendation on a species within this group that might be appropriate.

- **Node 6.2: Nematode genome diversity.** Like arthropods, nematodes are a phylum within the diverse Ecdysozoan supertaxon. It is clear that *C. elegans* (in nematode clade V) has secondarily lost many genes, including some Hox and ParaHox genes and some components of hh signaling. Basal nematodes (clades I, II) have a fuller complement of Hox genes, suggesting retention of a more ancestral genome arrangement. With the closely related *C. briggsae* genome, and the more distant *Brugia malayi* genome (clade III) (sequencing in progress), two other nematodes are good examples of candidates -- *Pristionchus pacificus, which is* intermediate in phylogenetic distance between *Caenorhabditis* and *Brugia*, and *Trichinella spiralis* from an early diverging lineage within nematodes. *Trichinella* is also a human and animal pathogen, enhancing its value as a candidate.

- **Node 6.3: Basal Ecdysozoa**. Current arthropod and nematode genome sequences are very divergent from one another, so there are big problems in deducing which genome characters are ancestral for the supertaxon of all ecdysozoans. This may become clearer as more arthropod and nematode genomes are sequenced. A candidate for future consideration is *Priapulus caudata* – a priapulid worm, perhaps a valuable basal sample. No basal ecdysozoan genome is included for sequencing in this current plan.

- **Part 4-Node 6.4: Lophotrochozoa.** This supertaxon of 15 phyla encompasses wide diversity. Three of the major phyla are molluscs, annelids and platyhelminths. We consider that

there should be one or more genome sequences for each. Current sequencing projects include *Schmidtea mediterranea* (a turbellarian platyhelminth in progress), *Schistosoma mansoni* (a trematode platyhelminth, the Schistosome parasite), and *Platynereis dumerilii* (a polychaete annelid approved to be done by Genoscope). Two mollusks deserve strong consideration: *Biomphalaria* sp. (*glabrata*) –a pulmonate gastropod mollusk, which has the added medical relevance of being the schistosome vector. In addition *Aplysia californica* should be considered as an opisthobranch gastropod mollusk used as a model system in neurobiology. While both species deserve high consideration, one may need to be prioritized over the other. If so, *Biomphalaria* would have higher priority because it is an important disease vector. In that case, the choice of a second mollusk genome would have to be considered again against other priorities.

- **Node 5**: **Acoelomorph** [Divergence in the Pre-Cambrian, >545Mya?]: This is a poorly studied group, but one of potentially of pivotal evolutionary importance as the sister group of all other bilateral animals. These animals have bilateral organization and a third (mesoderm) germ layer, but unlike most of the above bilateral animals, they have a blind gut and a diffuse nervous system. Unpublished studies indicate a smaller diversity of Hox genes than the Eubilateria, suggesting they may capture an intermediate stage in genome evolution between diploblasts and eubilaterians. One candidate would be *Convoluta pulchra* , which is a symbiont-free species of acoel. We do not propose an acoel at this time, but believe that more information should be sought to support this choice in the future.

- **Node 3/4 (node incompletely resolved)**: **Cnidarian and the radial diploblastic ancestor** [Deep pre-Cambrian, >560Mya?]: Cnidarians have neurons and contractile cells, epithelia with apical-basal polarized cells, two germ layers (classically no mesoderm, although this is contentious), a diffuse nervous system (perhaps two or three interpenetrating systems) with axon tracts, a cylindrical body plan (with rotational or biradial symmetry, or possibly cryptic bilaterality), and a blind gut. Despite the radically different body plan from bilateral animals, they have some signaling pathways comparable to those of bilaterians (e.g. complete Wnt pathway in *Hydra*). However, the diversity of developmental transcription factors is lower than in bilaterians; for example, fewer true Hox and ParaHox genes (with none in the central class), and Pax genes that cannot be readily assigned to the four bilateran classes. Complete genome information would resolve why some genetic characters are more stable than others, and would allow this to be related to cell-type evolution.

One species of cnidarian is scheduled for genome sequencing at JGI (*Nematostella vectensis*, an anthozoan). In view of the great diversity of this phylum (Anthozoa, Hydrozoa, Scyphozoa, Cubozoa) and the importance of this deep node, two other species should be considered: *Hydra vulgaris/magnipapillata* would be a desirable target because it is a hydrozoan and which also is used extensively for developmental studies. In addition, a species of coral should be considered, either an *Acropora* or a *Porites*, at a future time when enough information is available to prioritize between these two attractive candidates.

- **Node 3/4 (node incompletely resolved)**: **a ctenophore**: biradial body plan (cryptic bilaterality?), 15-20 cell types, diffuse nervous system with axon tracts, some mesoderm-like third layer with striated muscle. This node requires additional consideration before any specific recommendations can be made.

- **Node 1/2 (node incompletely resolved)**: **Sponges, at the base of the metazoa** [Divergence in the deep pre-Cambrian, >600Mya?]: a few cell types (5-6), basic multicellularity with cell adhesion, extracellular matrix (e.g., collagen), some cell-cell signaling (RTK receptors), perhaps epithelia, local organization but no true body plan. Further consideration of this important node is a high priority, but identification of appropriate species within the sponges could be a challenge because of their diversity and because practical issues (obtaining suitable DNA samples) may be important in making a choice.

- **Node 1/2 (node incompletely resolved) Placozoa, at the base of the Metazoa** [Divergence in the deep pre-Cambrian; >600Mya?]: Morphologically these are the most simple of metazoans, lacking organs or symmetry, possessing few cell types, with two cell layers and no gut. They have been shown, however, to possess clear orthologues of several developmental genes found in bilaterians and cnidarians, e.g. Cnox-2, Not, Brachyury. Placozoa possess a tiny genome (50Mb), possibly the smallest of any metazoan. The best candidate is *Trichoplax adhearens*.

C. What specifically do we expect this sequencing to reveal and why?

1. Proteome evolution: Many human genes (coding sequences) have originated within the Metazoa at various times during evolution and have been conserved through the human lineage to the present. Some genes are typical of animals, but have been lost in our own lineage. Those genes that encode proteins and RNAs involved in multicellular functions (as opposed to basic cell biological functions [see eukaryotic protists] or biochemical functions [see prokaryotes]) are most likely to have arisen within the Metazoa — for example, genes involved in morphology, physiology, behavior, and multicellular development, processes that are biological innovations of the Metazoa. As noted below, a number of them have been implicated in disease; indeed, one viewpoint is that the genes and

pathways underlying these innovations that lead to more complex anatomy and physiology are also the ones that are affected by disease in humans. It is expected that many of these proteins arose by exon shuffling or by duplication and diversification of pre-existing genomic sequences. Examples of such proteins that are expected to be identified within the Metazoa include:

- proteins of the 4 major kinds of cell-cell junctions, of components of the extracellular matrix, and of cell-cell adhesion (cadherins, CAMs, etc).
- protein components of the 17-20 major cell-cell signaling pathways (e.g. Wnt , hh, BMP, TGFβ, ephrins), numerous scaffolding proteins, and numerous signaling modifiers such as sprouty, bambi, hip, noggin, chordin, dickkopf, SFRPs, etc. These signaling pathways are altered in several kinds of cancer.
- many hundreds of kinds of protein kinases (ser-thr and tyr), and hundreds of kinds of E3 ubiquitin ligases, as regulatory components. Protein kinases have been implicated in cancer—the drugs Gleevec and Iressa specifically target protein kinases.
- proteins involved in the function of specialized cell types, such as various ion channels of nerve cells; oxygen carriers, myelination proteins of Schwann cells, muscle proteins (e.g., titin, myosin isoforms), various hormone and secretory products of gland cells. Defects in some of these proteins and the pathways in which they participate have been implicated in diseases such as multiple sclerosis and muscular dystrophy.
- great expansions and diversification of transcription factor families, especially those involved in the development of various cell types (e.g. such as myogenic factors, PPARgamma (implicated in diabetes), achaete-scute proteins) and of the development of compartments of the body plan, such as the ANTP class homeobox genes (including Hox genes, ParaHox genes, NK homeobox genes, En, Emx, Dlx etc.) and PRD class homeobox genes (including Pax genes, Otx, Gsc, Crx, Arix etc.)
- special cases of proteins such as keratin-type intermediate filament proteins arising in the chordates or noradrenalin production and reception in the vertebrates.

Once the sequences are available for comparison to human, it will be possible to trace human coding sequences backwards through the nodes and find their metazoan origins. Comparisons between *Drosophila* and mouse or human have been possible because once a sequence arises in evolution, its functional domains are extensively conserved thereafter in both lineages. Likewise, at earlier stages of protist evolution, before the Metazoa, many proteins of eukaryotic cell biology arose (e.g. cytoskeleton, mitosis and cell cycle, partitioning components to organelles) and were thereafter conserved. And before the protists, in the prokaryotes, many proteins of biosynthetic and energy metabolism, and translation arose and were thereafter conserved, up to humans.

2. Structural RNAs and small regulatory RNAs: A wide variety of genomic sequences are transcribed, not into pre-mRNAs destined for protein translation,

but into RNAs that have inherent function. The sequence conservation of these RNAs, and hence of the encoding DNA sequences, is extensive among animals, presumably because of the functional demands of complementary base-pairing within the RNA or between RNAs, or because of specific protein interactions. Major structural RNAs include the ribosomal 28S, 18S (this is used extensively in sequence-based phylogenies of animals), and 5S RNAs encoded in the DNA in long tandem repeats with particular spacings and transcriptional regulatory sequences, as well as the entire suite of tRNAs, the 7S RNA of the signal recognition particle (SRP), and various small nuclear RNAs such as those of the splicosome and of the small nucleolar RNAs.

Recently a large class of regulatory microRNAs (miRNAs) has been discovered., Approximately 22 nucleotides (+/-2) in length, these entities are thought to block translation of specific mRNAs by hybridizing to complementary short sequences in the 3' UTRs. The miRNAs are encoded in longer sequences that, when transcribed, form stem loop structures, from which a short dsRNA is cleaved. From this dsRNA, an anti-sense strand is kept. (Recent reviewed in (1): The *lin-3* and *let-7* miRNAs of *C. elegans* were the founder examples, but examples are now recognized in many animals, including humans (at least 145 kinds estimated), *Drosophila* (at least 71 kinds estimated) and *C. elegans* (at least 100 kinds estimated), as well as in plants. The *let-7* sequence is conserved across numerous phyla of bilateral animals. Some of these miRNAs are encoded in clusters. Other regulatory RNAs, the small interfering RNAs (siRNAs) provide specificity for certain endonucleases, leading to the actual degradation of mRNAs. Furthermore, several small modulatory dsRNAs (smRNAs) have also been discovered recently (e.g., in neural stem cells) and appear to be additional key, non-coding regulators of cell behavior at the transcriptional and posttranscriptional levels.

All of these small RNAs are thought to have regulatory roles in development and in specialized cell types, although at present the targets of their regulation have been established in only a few cases. Since small regulatory RNAs also operate in bacteria, it is likely this mode of regulation is very ancient. Comparative genomics could reveal the duplication, clustering, diversification, and loss of such important sequences.

3. Intron-exon organization of genes: From the composite nature of their encoded proteins, some human genes appear to have arisen by exon copying and shuffling within the Metazoa. Such proteins may contain 5-10 widely used domains (motifs), in different combinations and orders (e.g., in ECM proteins, blood clotting proteins, and many transmembrane receptors).

4. Gene family expansion and diversification: As noted above, many families of transcription factors have expanded and diversified within the Metazoa, including Fox (forkhead), bHLH, zinc-finger, homeodomain, T-box, nuclear receptor, etc. Interestingly, some of these are (or were) arranged in genomic clusters, perhaps

related to gene regulation and function. The Hox clusters are the best known, but recent data also point to ancient clustering of ParaHox, NK homeobox and some Fox-class transcription factors. Comparative genomics will demonstrate the ancestry and antiquity of these clusters, and reveal the patterns of expansion, conservation, and break-up in relation to animal diversification. Other unanticipated, ancestral genes linkages are almost certain to emerge from this program of comparative genomics, with implications for understanding the functional significance of gene order and position.

5. cis-Regulatory sequences: Because of the great evolutionary distances of the comparisons to be made among these genomes, it is likely that some cis-regulatory sequences will have changed too much to be recognizable as derived from a common ancestral sequence. But some cis-regulatory sequences may be conserved, and they would therefore be particularly noteworthy . Long-range gene clustering is also likely to be relevant for correct gene regulation, and this will be revealed by comparative genomics.

D. Quantitative expectations

Most of the comparisons of this section involve the genomes of species separated from humans by 500 million years or more. There is no doubt whatsoever that, at these long divergence times, all neutral positions will be completely mutation-saturated, i.e., the mean number of substitutions per position is >>1. Thus, when a statistically significant similarity between stretches of sequences is detected, there is no doubt that these sequences have been subject to strong purifying selection, i.e., comprise important signals. At the same time, while comparison of protein sequences at these evolutionary distances is usually straightforward, the comparison of non-coding sequences is not. When long nucleotide sequences, in which only small islands are expected to be conserved (as in cis-regulatory regions) are compared, great care is necessary to detect those correctly, without either missing the signals or erroneously taking noise for a signal. However, some recent methods seem to do a good job even under these difficult conditions, especially when enough sequences are available to do multiple comparisons (see, e.g. reference 2).Hence, it is important to have a tree reasonably densely populated with genomes in order to detect the most evolutionarily conserved regulatory signals.

With regard to protein evolution, it is useful to be clear about what exactly to expect. Even evolutionary distances of 500 MY are too short to produce much useful information about the evolution of sequences of highly conserved domains. In the case of fast evolving domains, particularly extracellular ones, the comparisons will be informative. However, the most important changes revealed by comparisons among the genomes of the set of organisms that this working group has addressed will actually be of a different kind. These comparisons will reveal changes of gene repertoire, including lineage-specific gene loss and expansion of paralogous families, and changes of domain architecture of

orthologous proteins, including so-called domain accretion. Remarkable diversity at these levels has already been revealed by comparison of the human, mouse, fly, and nematode proteomes (e.g., references 3,4). Genomes from lineages that branched off both before and after the arthropod-vertebrate divergence are critical for developing an adequate picture of these evolutionary processes.

Section 1 references

1. Bartel, D.P. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116: 281-97 (2004).
2. Roytberg MA, Ogurtsov AY, Shabalina SA, Kondrashov AS. A hierarchical approach to aligning collinear regions of genomes. Bioinformatics 18(12):1673-80 (2002).
3. Lespinet O, Wolf YI, Koonin EV, Aravind L. The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res. 12(7):1048-59 (2002).
4. Koonin et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. 5(2):R7(2004).

Section 2: Linking genomic change to life history and behavior

A. *The case for sequencing large genomes*

While there is significant advantage at present to the choice of smaller genomes for sequencing, much of animal phylogeny consists of taxa comprising organisms with large genomes. There are three reasons to sequence larger genomes when sequencing costs have been sufficiently lowered.

First, limiting the available genomic sequences to those of organisms with smaller genomes may lead to a biased view of genomes, since the small ones may tend to have reduced gene sets and/or a specialized organization and architecture. A clear example is the Hox genes, where the small genome of *C. elegans* has a reduced set of Hox genes, and the small genome of *D. melanogaster* has a disrupted organization of Hox genes.

Second, small genomes have a biased taxonomic distribution. For example, among the arthropods the holometabolous insects tend to have relatively small genomes, while most of the remaining insects (the many hemimetabolous members) and most of the crustaceans have genomes approximately an order of magnitude larger. As an example, grasshoppers have genomes of approximately the same size as the human genome, whereas fruit flies have a genome approximately 5% that of the human. Within the exdysozoans, no large genome has been sequenced and, therefore, the extent of bias in our understanding of genome architecture is unknown. This is potentially very relevant to the extrapolation of biological understanding from the protostome model systems (e.g., *C. elegans* and *D. melanogaster*, both of which have small genomes) to human (with a rather large genome) biology.

Finally, variation in genome sizes has long provided a great mystery for biologists. This is the so-called C-value paradox, the phenomenon that, even within a rather closely defined group of animals, such as insects or amphibia, genomes might vary 10-100 fold in size (e.g. many newts have genomes 20 times the size of the human genome, while some frogs have genomes only 0.2 times that of the human genome) even though the various species seem similar in complexity of anatomy, physiology, and behavior. The causes and correlates of this variation have been difficult to identify. It is not attributable to endopolyploidy. The difficulty in understanding variation in genome sizes may arise from our primitive understanding of the differences between large and small genomes. One way to approach this problem would be to compare large and small genome sequences from reasonably closely related organisms within taxa.

The major problem with proposing large genomes for sequencing at this time is cost, but that problem is likely to diminish with time. In the meantime, there may be efficient means to gather comparative genome data on large versus small genomes. Sequencing of targeted genomic regions, light shotgun sequencing, or sequencing of non-repetitive regions would all provide some data on comparative

genome architecture. Each of these options involves different tradeoffs, which must be discussed along with potential sequencing strategies, before useful examples can be proposed.

B. *The impact of asexuality on genome architecture.*

Sexuality is the predominant mode of reproduction among animals and other eukaryotic organisms. Although it has been difficult to argue the evolutionary advantages for the initial establishment of this mode, its persistence over long periods of evolution is generally seen in terms of the advantage of the increased genetic variation it provides within populations of animals. Does sexuality require certain traits of genome architecture? This question might be answered by comparing genomic sequence information from known lines of bdelloid rotifers that have probably remained asexual for many millions of years, with the sequence of the genomes of closely related lines that are sexual. The high level of interest in the question is balanced by the large size of the relevant rotifer genomes. The current plan does not include sequencing of these interesting genomes because of the large size of the genomes.

Section 3: Protist origins of the human genome.

A. Background

Microbes were the only form of life for the first 2-3 billion years of planetary and biological evolution. The first eukaryotes were single-cell creatures that inhabited the earth at least 2.2 billion years ago, whereas the origins of multi-cellular plants, animals and fungi trace back no more than one billion years ago. Paleontological studies reveal that protists most closely resemble early eukaryotic life forms. Morphological descriptions and molecular sequence data show that the diversity of protists dwarfs that seen for the combined world of plants, animals and fungi. Protists represent an eclectic group of organisms marked by numerous innovations in body plans, ecology, biochemistry, and molecular processes. There are approximately 85-120 distinct protistan lineages that describe ~30,000 species, mostly unicellular life forms (http://www.mbl.edu/microscope). Free-living protists thrive in diverse environments but some of these are opportunistic human pathogens i.e. *Acanthamoeba*, *Naegleria* and *Balamuthia*. Other species are obligate parasites of animals, plants, fungi and other protists [1, 2].

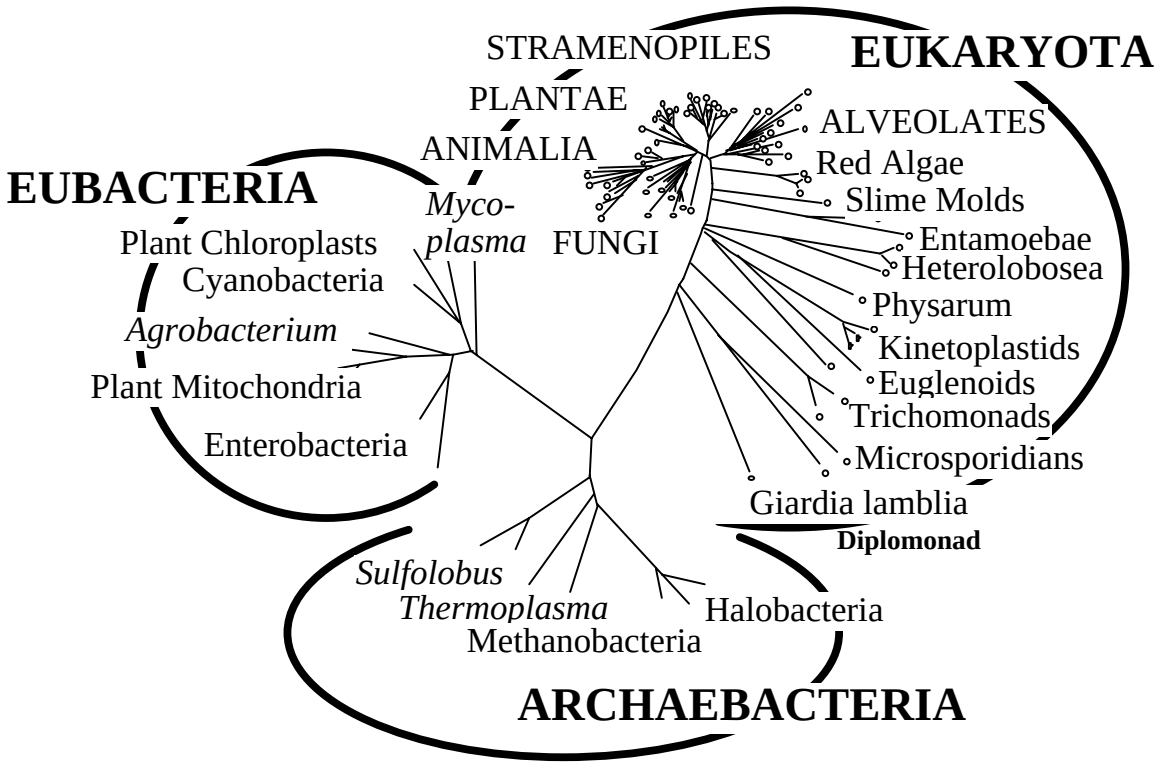


Figure 2. Universal Phylogenetic Tree based upon analyses of rRNA sequences. Structural similarities for sites that can be unambiguously aligned for these 100 representative taxa (approximately 1200 sites) were converted to evolutionary distances [4]. We employed "minimum evolution" [5] to infer a tree in which the length of line segments represents the evolutionary distance between taxa. Diplomonads, trichomonads and microsporidia are basal to all other eukaryotes, but phylogenies based upon comparisons of alternative gene families indicate that microsporidia are misplaced in rRNA trees [6-8].

Early phylogenetic inferences based upon comparisons of ribosomal RNA sequences [3], revolutionized perceptions of microbial relationships. The rRNA tree in **Figure 2** suggests that diplomonads, trichomonads and microsporidia were ancestral to all other eukaryotes [9-11]. These organisms lack mitochondria, peroxisomes and have trans-Golgi networks rather than stacked dictyosomes. This implies that eukaryotes may be older than once thought possible and, given the amitochondriate phenotype of early-diverging lineages, that the presence of oxygen is not prerequisite to forming a nucleus. In **Figure 2**, a series of independently branching lineages followed the early evolution of amitochondriate taxa with plants, animals and fungi emerging from a sudden evolutionary radiation that occurred no more than one billion years ago. Phylogenetic analyses of rRNAs have allowed the positioning of many taxa of uncertain affiliation, e.g., *Pneumocystis* as a member of the fungi [12] and *Cyclospora* as a close relative of *Eimeria* species [13]. These molecular trees also describe new complex evolutionary assemblages [12, 14-16] or confirmed relationships previously inferred from comparisons of morphology and ultrastructure synapomorphies [17].

When measured in terms of lifestyle and phenotypic variation, some of these major protist clades appear to be as complex as the traditional kingdoms Viridiplantae, Animalia and Fungi. Examples include the Alveolata (ciliates, dinoflagellates, and apicomplexans), the Stramenopiles (most golden brown algae, diatoms, yellow algae, oomycetes, labyrinthulomycetes, numerous heterotrophic flagellates, etc.) and the Opisthokonta (animals, fungi, choanoflagellates, nucleariids, and mesomycetozoans). Membership in each of these complex assemblages often supports ultrastructure and/or other phenotypic synapomorphies. Yet, many uncertainties remain. Due to abrupt radiations in the eukaryotic framework, there is little resolution among some of the major protist clades in rRNA-based comparisons. More importantly, we still do not have a clear understanding about which extant protist lineages might represent the earliest branches in the eukaryotic line of descent [11, 18]. Perhaps stochastic errors in single-gene phylogenies account for unresolved radiations or even the misidentification of basal branches [18, 19]. Such artifacts are bound to occur when the phylogenetic reconstruction algorithms do not adequately model the complexity of sequence evolution for a particular gene. To address this problem and to test the rRNA scheme, several groups shifted their attention to phylogenetic analyses of other gene families [20-25]. Combined phylogenies of multiple gene families have identified potential superclades of protists, e.g. a relationship between alveolates and heterokont protists, and the Euglenozoa with the heteroloboseans [26]. In some cases multigene phylogenies confirmed what was learned from rRNA comparisons, while in others, disagreements suggest alternative scenarios for the evolution of protists [26-28].

Despite these uncertainties, there is general agreement that protists are far more diverse than plants, animals and fungi. Clearly the transition from prokaryotic life forms to eukaryotic cellular architecture, and the origins of greater biological complexity as represented by multicellular plants, animals and fungi are rooted in the world of protists. Within the last few years, the broader scale application of genomic science has led to a renaissance in the study of microbial molecular evolution, especially for the Archaea and the Bacteria. There are more than two hundred prokaryotic genomes in public databases (http://www.tigr.org, http://ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html) with many more at different stages of completion (http://www.tigr.org/tdb/mdb/mdbinprogress.html). Despite the relatively impoverished state of morphological analysis for prokaryotes, in the near future detailed information about the evolution of their genomes will eclipse what we know about the evolution of morphologically rich organisms, such as plants, animals and fungi. The remaining gap in our knowledge of life's history on Earth will reside in the uncertain relationships of divergent eukaryotic microorganisms.

The comparison of genome sequences from phylogenetically divergent protists offers much more than a reference point for estimating potential genetic diversity

of the genes common to animals, plants or fungi. The choice of protist genome projects can be targeted to address higher-level scientific questions including: *What is the composition and relative branching order of the major eukaryotic lineages? To what extent has lateral gene transfer affected the evolution of eukaryotic genomes? Is there a genomic core that is refractory to horizontal gene transfer? Does the pattern of metabolic evolution mirror environmental changes in earth's history? Does adaptation to parasitic life-styles impose major changes in genome architecture and the transcriptome? What mechanisms explain the emergence of parasitism? What caused the elaboration of complex gene families and genome organization patterns that led to modern eukaryotic complexity as represented by animals, fungi, and plants and other recently derived groups?*

The available models for understanding how complexity evolved in eukaryotes are in flux. For example, interpretations of early molecular trees argued that the most basal protist lineages lacked introns and mitochondria [29, 30]. However, the discovery of proteobacterial-like molecular chaperones in all of the putatively deep branching amitochondriate protists [31-34] suggests that symbionts ancestral to mitochondria and hydrogenosomes could have been present in the early stages of eukaryote evolution [35]. Thus, although these anaerobic early-branching organisms were thought to reflect the ancestral condition of eukaryotes, it is becoming clear that some of their features (e.g. lack of mitochondria) may be secondary adaptations to anaerobic environments. These discoveries emphasize the fact that our molecular perspective of early eukaryote evolution is strongly biased by the selection of taxa and gene families that are currently available for molecular studies. There are many other amitochondriate flagellates that lack dictyosomes and complex cytoskeletal systems, but few of these lineages have been included in molecular trees. These protists occur in anoxic environments [20] and are only now being cultured and characterized in detail at the ultrastructural level [36]. Since some are likely to represent even deeper diverging eukaryote lineages, their phenotypic properties will be important in understanding the evolution and assembly of the first eukaryotes.

Broad-scale genomic sampling from taxa that might represent basal lineages in the eukaryotic line of descent is very limited. The *Giardia lamblia* (diplomonad) genome project (http://www.mbl.edu/Giardia) [37] provides the first glimpse of a genome from a putative, early diverging, amitochondriate protist. Certain features of the *Giardia* genome resemble prokaryotes, including strong similarities to proteins that serve metabolic functions, dense organization of protein coding regions oftentimes separated by only a few base pairs, a complete absence of myosin, and a general paucity of introns (although they possess spliceosomal machinery and at least one example of a spliceosomal intron [38]). There are no reports of mitochondria or derived organelles in *Giardia,* but its genome contains a mitochondrial-like cpn60 gene [34] and a mitochondrial type iron-sulfur cluster protein (IscS) [39]. This tentatively

suggests that a mitochondrion present when *Giardia* diverged from other eukaryotes was discarded as *Giardia* adapted to a microaerophilic niche. However, based upon complete information from the *Giardia* genome, there are no other credible examples of coding regions derived from the symbiont that would have been ancestral to mitochondria (Sogin and Morrison, manuscript in preparation). In other regards, the *Giardia* genome resembles more commonly studied eukaryotes, albeit with highly diverged coding regions. On the other hand, systems for transcription, DNA replication, translation, and polyadenylation in Giardia have fewer protein components than do such systems in yeast, plants, animals and most other protists.

In addition to limited taxon representation, alternative mechanisms of genome evolution might explain discrepancies between molecular trees inferred from different protist gene families. Paraphyly of ancestral gene duplications or horizontal gene transfer mediated by formation of chimerical genomes, endosymbioses, or viral-like mechanisms might be responsible for conflicting phylogenies. Single-gene phylogenies cannot resolve which of these mechanisms might be major factors in the evolution of eukaryotic genomes. To understand phylogenetic patterns for protists and their pivotal position in the evolution of complex life forms, we must adopt a strategy that includes and combines larger molecular and non-molecular datasets from a greater collection of taxa. The most aggressive strategy would be to sequence the genomes of all species that might inform us about protist evolution, in an attempt to understand how major eukaryotic innovations including ultrastructure, morphology (body plans), and life styles contributed to the evolution of the deepest branching eukaryotes, as well as lineages that were ancestral to the multicellular world. Such an enterprise, however, would require resources that eclipse the cost of the human genome project by several orders of magnitude and so is impractical. Exclusive of the Fungi, eukaryotic microbial genome projects are, at present, limited to a handful of free living protists (*Dictyostelium, Tetrahymena, Paramecium, Phytophthora, Thalassiosira*) and an incomplete sampling of parasitic protists (*Entamoeba, Giardia, Trypanosoma, Leishmania, Trichomonas, Encephalitozoon, Enterocytozoon, Nosema* plus 14 public and private

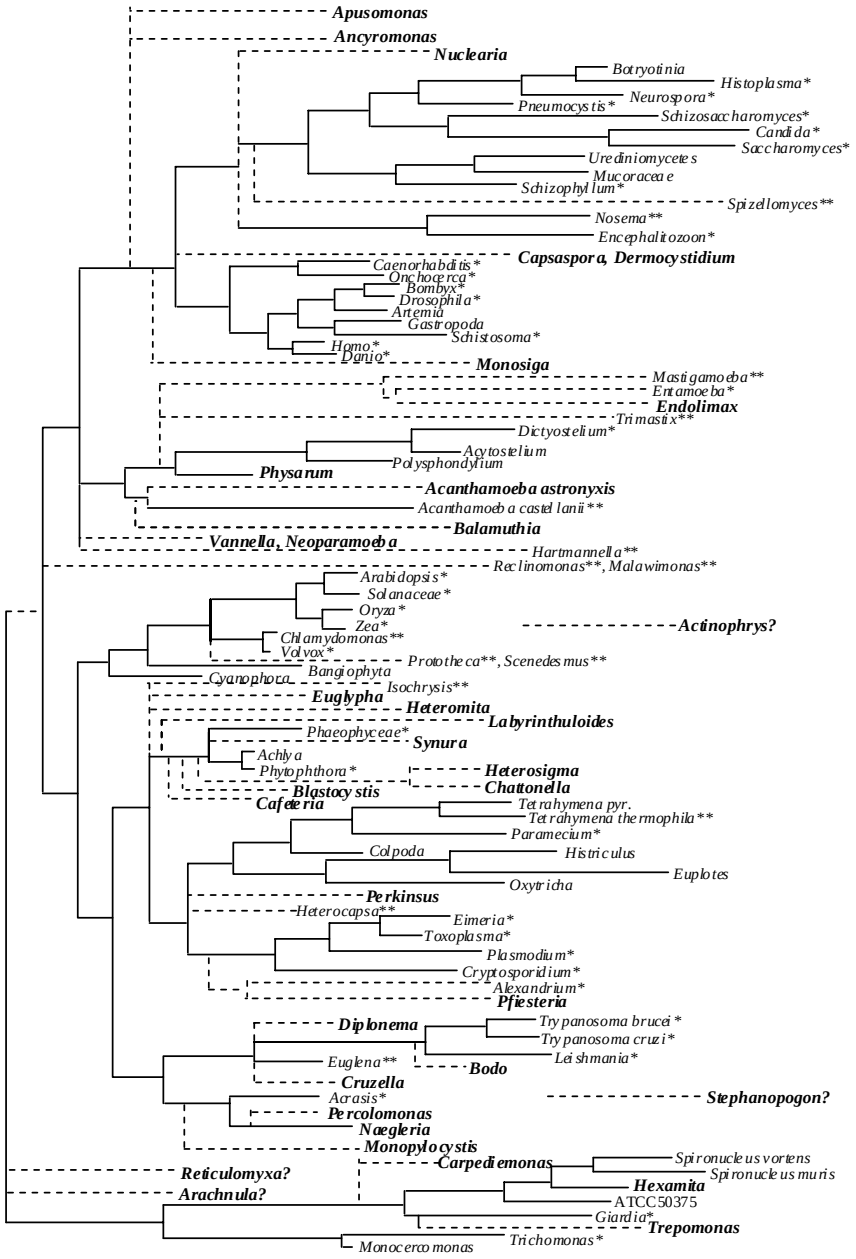


Figure 3. A widely cited phylogenetic tree based upon comparisons of a handful of distinct coding regions [13]. Although this phylogenetic analysis does not include all of the known protist assemblages, it does provide a framework for displaying the phylogenetic diversity of protist cDNA and genome projects that are currently in progress (indicated by * or **(PEP-Protist EST Project)) as well as candidate taxa for genome sequencing (bold lettering). Lineages with dashed lines are grafted onto Baldauf et al.'s original phylogeny at locations predicted by rRNA phylogenies and "?" indicates unknown phylogenetic positioning of a lineage.

Apicomplexa genome projects at various stages of completion– mostly *Plasmodium* species but also *Cryptosporidium, Toxoplasma, Eimeria and Theileria*). Some lineages are oversampled while others, such as choanoflagellates, which have the potential to inform us about the more recent origins of animals, fungi, and plants have been completely ignored. Other

important protists, including human parasites e.g. *Acanthamoeba*, *Naegleria*, etc. have not yet been selected for genome or cDNA analysis.

In a pioneering effort to survey the evolutionary breadth of protists, a consortium of Canadian investigators has recently initiated a large-scale cDNA (EST) sequencing project from a wide diversity of protists. The description on their website is as follows:

> The Protist EST Program (PEP) is a pan-Canadian collaboration, whose objective is to determine the expressed portions of genomes from a taxonomically broad collection of mostly unicellular eukaryotes. Representatives of ~25 different protist groups is being targeted for sequencing (either ESTs or complete cDNAs). All together upwards of 30-40 protists will be sampled, focusing in the initial phase on about 20 species (see http://amoebidia.bcm.umontreal.ca/pepdb/pep.php for listing). The information emerging from this undertaking will be used to address various issues relating to eukaryotic diversity, classification, evolution, and the origin of the eukaryotic cell.
>
> PEP is a large-scale interdisciplinary, and collaborative research project, involving six Canadian universities in five provinces. It is financed by Genome-Canada and managed by Genome-Atlantic and Génome Québec. PEP aims at the exploration of the diversity of eukaryotic genomes in a systematic, comprehensive and integrated way. The focus is on unicellular microbial eukaryotes, known as protists. Protistan eukaryotes comprise more than a dozen major lineages that, together, encompass more evolutionary, ecological and probably biochemical diversity than the multicellular kingdoms of animals, plants and fungi combined. PEP is a unique endeavor in that it is the first phylogenetically broad genomic investigation of protists.

Although the PEP's choice of organisms and the resulting EST data will be of enormous help in deciding on species for genome projects, the sequences are not yet publicly available. Thus, for the current proposal, a few important protists have been chosen for which many types of data already exist to support their candidacy for genome sequencing.

B. Priority Recommendations:

Each of the estimated 80 or more distinct protist lineages has a story to tell about evolutionary trajectories for eukaryotic cells. In making informed decisions about which of those histories and their associated biology are most important to recapitulate through the lens of genome science, key issues include, but are not limited to:

- *exploring genome architecture for basal branches,*

- *identifying the genomic influence of organelle acquisitions through symbiotic events, and*
- *identifying features of protist genomes that were immediately ancestral to animals plus fungi.*

There are so many excellent candidates for genome investigations that it is impossible to address all germane questions. Some general criteria that were used in making the recommendations include *phylogenetic position, genome size, medical relevance, and community interest*, but other scientific and practicalcriteria are also meaningful.

Figure 3 employs a multi-gene phylogeny to illustrate the diversity of protists. The working group recommended that the effort initially focus on a handful of taxa that are tractable for library construction, have or will soon have a set of ESTs available, and are maximally informative about major eukaryotic evolutionary innovations. Genera that satisfy these conditions include the following:

Oxytricha
Monosiga
Physarum
Symbiodinium
Acanthamoeba
Naeglaria
Reclinomonas
Endolimax
Retortamonas
Spironucleus
Carpediemonas

The first three taxa are probably the best candidates for genome sequencing at this time.

a) The spirotrichous ciliate *Oxytricha trifallax.* On the basis of an earlier white paper (http://www.genome.gov/10002154). , *O. trifallax* has previously been assigned high priority by NHGRI and the working group confirms that recommendation. The *Oxytricha* sequence will inform the 108 Mb macronuclear genome sequence of the distantly related *Tetrahymena* which is currently being completely sequenced. At 50 Mb, the *Oxytricha* macronuclear genome is small, but its “nanochromosomes” (average size 2 kb) require will modification to existing sequencing strategies.

b) The choanoflagellate *Monosiga ovata* represents a lineage that has long been thought to be at the base of the animal/protist divergence. Thus its sequence would inform the understanding of metazoan origins. While a

congener, *M. brevicolis,* is currently in the sequencing pipeline at JGI, these two species are not highly related, and the shared genus may even be a misnomer. Furthermore, *M. ovata* is thought to be more basal and *M. brevicolis* more derived. Pilot sequencing of *M. ovata* will be necessary to assess whether bacterial contamination will be a barrier to obtaining high coverage. As an important resource, roughly 11,500 ESTs for *M. ovata* have been generated in labs in Japan and England, and it has a small diploid genome size of 39 Mb.

c) *Physarum polycephalum*, an amoebozoan and myxomycete (acellular slime mold), represents an extremely poorly surveyed portion of the tree (only the Dictyostelium genome has been sequenced). Furthermore *Physarum* has been an experimental biology system for over 50 years. To give just one example of the reasons this species is a good choice for genome sequencing, it has naturally and perfectly synchronous populations of nuclei within a single plasmodium. This allows the detection of fine changes in chromatin structure that correlate with activity. Its mitochondria have a complex RNA editing system involving more types of events than any other gene editing system, and the biochemical components that mediate *Physarum* RNA editing all appear to be nuclearly encoded. Recently, post-genomic techniques such as RNAi have been extended to *Physarum*. A *Physarum* sequence would complement the *Dictyostelium* sequence, although at 14.5 % protein divergence, this is comparable to the human-*Drosophila* split. The genome size (300 Mb) is modest and *Physarum* has a molecular biology community poised to take advantage of the data. Currently, 3500 ESTs have been sequenced by PEP and a total of 10,000 are planned.

In addition to these three, the working group recommended that initial sample sequencing of representatives of the remaining set of 8 taxa would provide preliminary data to make an informed choice among them for additional genomic sequencing targets. Scientific justification for each of these is included below. In the case of *Symbiodinium*, the recommendation for pilot sequencing was based on the very large size of the genomes of dinoflagellates, this one at 2 Gb being one of the smallest. It is anticipated that a BAC library may become available; if so, it would be useful to sequence a few BACs to assess gene density, among other features of the genome. For the remaining 7 protist taxa, pilot sequencing is needed to assess feasibility. It would be reasonable to sequence both ends of 100 or 96 clones to test for bacterial or other food contamination, and if this does not pose a problem then to continue with another set of paired-end sequences of 1000 clones or more. Libraries are already available in many cases

d) The dinoflagellate *Symbiodinium.* The scientific case for sequencing this genome is strong, but the genome size is large for a protist (2 Gb). However, *Symbiodinium* may well turn out to be the smallest genome, and hence most tractable, among the dinoflagellates, which are the sister group to the parasitic apicomplexans and ciliates. Because of the cost of

full sequencing, pilot sequencing should be pursued, particularly if BACs become available to estimate gene density.

e) *Acanthamoeba* is a strong candidate for pilot sequencing to test feasibility. In addition to its phylogenetic importance, *Acanthamoeba* is a human pathogen. An EST collection exists for this organism.

f) The heterolobosean (amoeboflagellate) *Naegleria*. The major motivations for choosing *Naegleria* [41] are that: i) it represents a potentially deep-branching lineage of eukaryotes for which no genome project is currently underway (although it may be a very distant relative of the trypanosomes), ii) it is an organism of interest to cell biologists, in that it transforms from an amoebal to flagellate stage [42], and iii) species include both free-living and pathogenic forms, allowing the direct comparison between them [43]. Three species should be considered: *N. gruberi* (a laboratory species for which cell biological studies have been done), and *N. fowleri* (a pathogenic species that causes encephalitis), *N. lovaniensis* (a free-living strain, evolutionarily intermediate between *N. gruberi* and *N. fowleri [44]*). The genome size estimates [45] range from 19-104 Mb depending on method (sum of chromosome bands *vs*. direct c-value measurements, respectively). There are some existing resources for genomic and cDNA libraries, including a pilot GSS and EST project (A. Roger, above).

g) The jakobid flagellate, *Reclinomonas americana* (and/or another tractable jakobid relative). *Reclinomonas* is a member of diverse assemblage of protists that likely fill an important evolutionary place in eukaryotic history. Specifically, *Reclinomonas* and jakobid relatives have by far the most complex, gene-rich mitochondrial genome known among eukaryotes [46]. This alone positions jakobids among the deepest-branching of protists, making them particularly important for evolutionary reasons. [47] Unfortunately, however, very little is known about their genomic biology. In a recent pilot project by Franz Lang (U. Montreal, Canada) and colleagues, genome size estimates are being carried out as is some pilot GSS sequencing. In addition, ~10,000 ESTs have already been generated for *Reclinomonas* and ~5000 from *Malawimonas* from the PEP project.

h) The diplomonad *Spironucleus barkhanus* (ATCC 50380). The main reasons for choosing this organism, an important parasite of fish, are that: i) it is a relative of *Giardia lamblia*, whose genome has been recently completed; as such, its sequence will be greatly beneficial in annotating and understanding the *Giardia* genome. ii) it likely has a similarly small genome to *Giardia (*see below) and, 3) it (along with *Giardia*) represents a potentially deep-branching lineage in the eukaryotic tree [40]; thus, it will serve to enhance our understanding of the origin and evolution of eukaryotic cells. Both cDNA and genomic libraries have already been prepared from *S. barkhanus,* and a small insert library (1-5 kb) is available

for pilot sequencing. Indeed, a project initiated by Andrew Roger (Dalhousie University, Canada) and colleagues has generated ~6200 Genome Survey Sequences (GSS) and ~2500 ESTs from these libraries, but the data are not publicly available. Although a precise genome size estimate is not yet available for *Spironucleus,* these pilot data indicate a genome of similar composition and gene density as that of *Giardia* (~12Mb).

i) *Endolimax,* or a pelobiont at the base of the entamoebae, would inform relationships to the parasitic entamoebae. It is either basal to other entamoebae or it is a sister group. Entamoebae are significant pathogens causing high levels of morbidity and mortality in developing countries. Both lack mitochondria and a microtubule-based cytoskeleton and flagella. These are thought to be secondary losses. The *Entamoeba* genome is at a draft stage, and the *Endolimax* genome could be of value for interpreting it.

j) *Retortamonas,* an amitochondriate protist, would be a candidate for analyzing the deep divergence among diplomonads.

k) *Carpediemonas,* another amitochondriate protist, has small micro bodies that have been interpreted as relic mitochondria, but this is unresolved.

The last two candidates, *Retortamonas* and *Carpediemonas*, are both difficult to grow, and the arguments are similar for selecting either, so pilot sequencing would be reasonable to assess feasibility of either one for further studies, and to gather preliminary data to make an informed recommendation.

Section 3 references:

1. Patterson, D.J., *The Diversity of Eukaryotes.* American Naturalist, 1999. **154**: p. 96-124.
2. Margulis, L., et al., *Handbook of the Protoctista*, ed. H.I. McKhann. 1990, Boston: Jones and Bartlett Publishers. 914 pp.
3. Woese, C.R., *Bacterial evolution.* Microbiol Rev, 1987. **51**(2): p. 221-71.
4. Elwood, H.J., G.J. Olsen, and M.L. Sogin, *The small-subunit ribosomal RNA gene sequences from the hypotrichous ciliates Oxytricha nova and Stylonychia pustulata.* Mol Biol Evol, 1985. **2**(5): p. 399-410.
5. Rzhetsky, A. and M. Nei, *A simple method for estimating and testing minimum-evolution trees.* Mol. Biol. Evol., 1992. **9**: p. 945-967.
6. Edlind, T.D., *Phylogenetics of protozoan tubulin with reference to the amitochondriate eukarytoes*, in *Evolutionary Relationships Among Protozoa*, G.H. Coombs, et al., Editors. 1998, Kluwer Academic Publishers: Dodrecht. p. 91-108.
7. Keeling, P.J., M.A. Luker, and J.D. Palmer, *Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi.* Mol Biol Evol, 2000. **17**(1): p. 23-31.
8. Hirt, R.P., et al., *Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins.* Proc Natl Acad Sci U S A, 1999. **96**(2): p. 580-5.
9. Vossbrinck, C.R., et al., *Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes.* Nature, 1987. **326**(6111): p. 411-4.
10. Sogin, M.L., et al., *Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from Giardia lamblia.* Science, 1989. **243**(4887): p. 75-7.
11. Leipe, D.D., et al., *Small subunit ribosomal RNA+ of Hexamita inflata and the quest for the first branch in the eukaryotic tree.* Mol Biochem Parasitol, 1993. **59**(1): p. 41-8.
12. Edman, J.C., et al., *Ribosomal RNA sequences show Pneumocystis carinii to be closely related to the yeasts.* Nature, 1988. **334**: p. 519-522.
13. Relman, D.A., et al., *Molecular phylogenetic analysis of Cyclospora, the human intestinal pathogen, suggests that it is closely related to Eimeria species.* J Infect Dis, 1996. **173**(2): p. 440-5.
14. Gajadhar, A.A., et al., *Ribosomal RNA sequences of Sarcocystis muris, Theileria annulata and Crypthecodinium cohnii reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates.* Mol Biochem Parasitol, 1991. **45**(1): p. 147-54.
15. Sogin, M.L. and J.D. Silberman, *Evolution of the protists and protistan parasites from the perspective of molecular systematics.* Int J Parasitol, 1998. **28**(1): p. 11-20.
16. Wainright, P.O., et al., *Monophyletic origins of the metazoa: an evolutionary link with fungi.* Science, 1993. **260**(5106): p. 340-2.

17. Gunderson, J.H., et al., *Phylogenetic relationships between chlorophytes, chrysophytes, and oomycetes.* Proc Natl Acad Sci U S A, 1987. **84**(16): p. 5823-7.
18. Philippe, H., A. Germot, and D. Moreira, *The new phylogeny of eukaryotes.* Curr Opin Genet Dev, 2000. **10**(6): p. 596-601.
19. Philippe, H. and A. Germot, *Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution.* Mol Biol Evol, 2000. **17**(5): p. 830-4.
20. Edgcomb, V.P., et al., *Benthic eukaryotic dirversity in the Guaymas Basin hydrothermal vent environment.* Proc. Natl. Acad. Sci, 2002 in press. **in press**.
21. Edlind, T., et al., *Cryptosporidium and microsporidial B-Tubulin sequences: prediction of Benzimidazole sensitivity and phylogeny.* J. Eukaryotic Microbiol., 1984. **41**: p. 385.
22. Baldauf, S.L. and J.D. Palmer, *Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins.* Proc. Nat'l. Acad. Sci. U. S. A., 1993. **90**(24): p. 11558-62.
23. Baldauf, S.L., J.D. Palmer, and W.F. Doolittle, *The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny.* Proc Natl Acad Sci U S A, 1996. **93**(15): p. 7749-54.
24. Hashimoto, T., et al., *Early evolution of eukaryotes inferred from protein phylogenies of translation elongation factors 1α and 2.* Archiv für Protistenkd, 1997. **148**: p. 287-295.
25. Keeling, P.J., N.M. Fast, and G.I. McFadden, *Evolutionary relationship between translation initiation factor eIF- 2gamma and selenocysteine-specific elongation factor SELB: change of function in translation factors.* J Mol Evol, 1998. **47**(6): p. 649-55.
26. Baldauf, S.L., et al., *A kingdom-level phylogeny of eukaryotes based on combined protein data.* Science, 2000. **290**(5493): p. 972-7.
27. Baldauf, S.L. and W.F. Doolittle, *Origin and evolution of the slime molds (Mycetozoa).* Proc Natl Acad Sci U S A, 1997. **94**(22): p. 12007-12.
28. Embley, T.M. and R.P. Hirt, *Early branching eukaryotes?* Curr Opin Genet Dev, 1998. **8**(6): p. 624-9.
29. Patterson, D.J. and M.L. Sogin, *Eukaryote origins and protistan diversity*, in *The Origin and Evolution of Prokaryotic and Eukaryotic Cells.*, H. Hartman and K. Matsun, Editors. 1993, World Scientific Publishing Co.: New Jersey. p. 13-46.
30. Cavalier-Smith, T., *Archamoebae: the ancestral eukaryotes?* Biosystems, 1991. **25**: p. 25-38.
31. Hirt, R.P., et al., *A mitochondrial Hsp70 orthologue in Vairimorpha necatrix: molecular evidence that microsporidia once contained mitochondria.* Curr Biol, 1997. **7**(12): p. 995-8.
32. Bui, E.T., P.J. Bradley, and P.J. Johnson, *A common evolutionary origin for mitochondria and hydrogenosomes.* Proc Natl Acad Sci U S A, 1996. **93**(18): p. 9651-6.

33. Peyretaillade, E., et al., *Microsporidia, amitochondrial protists, possess a 70-kDa heat shock protein gene of mitochondrial evolutionary origin.* Mol Biol Evol, 1998. **15**(6): p. 683-9.
34. Roger, A.J., et al., *A mitochondrial-like chaperonin 60 gene in Giardia lamblia: evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria.* Proc Natl Acad Sci U S A, 1998. **95**(1): p. 229-34.
35. Martin, W. and M. Müller, *The hydrogen hypothesis for the first eukaryote.* Nature, 1998. **392**: p. 37-41.
36. Walker, G., et al., *Ultrastructural identities of Mastigamoeba punctachora, Mastigamoeba simplex, and Mastigella commutans and assessment of hypotheses of relatedness of the pelobionts (Protista).* European Journal of Protistology, 2001. **37**: p. 25-49.
37. McArthur, A.G., et al., *The Giardia genome project database.* FEMS Microbiol Lett, 2000. **189**(2): p. 271-3.
38. Nixon, J.E., et al., *A spliceosomal intron in Giardia lamblia.* Proc Natl Acad Sci U S A, 2002. **99**(6): p. 3701-5.
39. Tachezy, J., L.B. Sanchez, and M. Muller, *Mitochondrial type iron-sulfur cluster assembly in the amitochondriate eukaryotes Trichomonas vaginalis and Giardia intestinalis, as indicated by the phylogeny of IscS.* Mol Biol Evol, 2001. **18**(10): p. 1919-28.
40. Rozario, C., et al., *Primary structure and phylogenetic relationships of glyceraldehyde-3-phosphate dehydrogenase genes of free-living and parasitic diplomonad flagellates.* Journal of Eukaryotic Microbiology, 1996. **43**(4): p. 330-340.
41. De Jonckheere, J.F., *A century of research on the amoeboflagellate genus Naegleria.* Acta Protozoologica, 2002. **41**(4): p. 309-342.
42. Fulton, C., *Naegleria - a Research Partner for Cell and Developmental Biology.* Journal of Eukaryotic Microbiology, 1993. **40**(4): p. 520-532.
43. Marciano-Cabral, F., *Biology of Naegleria spp.* Microbiol Rev, 1988. **52**(1): p. 114-33.
44. Zhou, L., et al., *Genetic variations in the internal transcribed spacer and mitochondrial small subunit rRNA gene of Naegleria spp.* J Eukaryot Microbiol, 2003. **50 Suppl**: p. 522-6.
45. Clark, C.G., *Genome structure and evolution of Naegleria and its relatives.* J Protozool, 1990. **37**(4): p. 2S-6S.
46. Lang, B.F., et al., *An ancestral mitochondrial DNA resembling a eubacterial genome in miniature.* Nature, 1997. **387**(6632): p. 493-7.
47. Lang, B.F., et al., *A comparative genomics approach to the evolution of eukaryotes and their mitochondria.* J Eukaryot Microbiol, 1999. **46**(4): p. 320-6.