2003

Gene discovery in the apicomplexa as revealed by EST sequencing and assembly of a comparative gene database

Deana Pape
Washington University School of Medicine in St. Louis

John Martin
Washington University School of Medicine in St. Louis

Todd Wylie
Washington University School of Medicine in St. Louis

Mike Dante
Washington University School of Medicine in St. Louis

Robert H. Waterston
Washington University School of Medicine in St. Louis

See next page for additional authors

Recommended Citation

Pape, Deana; Martin, John; Wylie, Todd; Dante, Mike; Waterston, Robert H.; Clifton, Sandra W.; and et al, ,"Gene discovery in the apicomplexa as revealed by EST sequencing and assembly of a comparative gene database." Genome Research. 13,. 443-454. (2003).
https://digitalcommons.wustl.edu/open_access_pubs/2111

Authors
Deana Pape, John Martin, Todd Wylie, Mike Dante, Robert H. Waterston, Sandra W. Clifton, and et al

Gene Discovery in the Apicomplexa as Revealed by EST Sequencing and Assembly of a Comparative Gene Database

Li Li, Brian P. Brunk, Jessica C. Kissinger, et al.

Genome Res. 2003 13: 443-454
Access the most recent version at doi:10.1101/gr.693203

Gene Discovery in the Apicomplexa as Revealed by EST Sequencing and Assembly of a Comparative Gene Database

Li Li,[1] Brian P. Brunk,[2] Jessica C. Kissinger,[1,3] Deana Pape,[4] Keliang Tang,[5] Robert H. Cole,[5] John Martin,[4] Todd Wylie,[4] Mike Dante,[4] Steven J. Fogarty,[5] Daniel K. Howe,[6] Paul Liberator,[7] Carmen Diaz,[7] Jennifer Anderson,[7] Michael White,[8] Maria E. Jerome,[8] Emily A. Johnson,[8] Jay A. Radke,[8] Christian J. Stoeckert Jr.,[2] Robert H. Waterston,[4] Sandra W. Clifton,[4] David S. Roos,[1] and L. David Sibley[5,9]

[1]Department of Biology, [2]Center for Bioinformatics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; [3]Department of Genetics, University of Georgia, Athens, Georgia 30602, USA; [4]Genome Sequencing Center, Department of Genetics, [5]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, Missouri 63108, USA; [6]Department of Veterinary Sciences, University of Kentucky, Lexington, Kentucky 40546, USA; [7]Human and Animal Infectious Diseases, Merck Research Laboratories, Rahway, New Jersey 07065, USA; [8]Veterinary Molecular Biology, Montana State University, Bozeman, Montana 59717, USA

Large-scale EST sequencing projects for several important parasites within the phylum Apicomplexa were undertaken for the purpose of gene discovery. Included were several parasites of medical importance (*Plasmodium falciparum*, *Toxoplasma gondii*) and others of veterinary importance (*Eimeria tenella*, *Sarcocystis neurona*, and *Neospora caninum*). A total of 55,192 ESTs, deposited into dbEST/GenBank, were included in the analyses. The resulting sequences have been clustered into nonredundant gene assemblies and deposited into a relational database that supports a variety of sequence and text searches. This database has been used to compare the gene assemblies using BLAST similarity comparisons to the public protein databases to identify putative genes. Of these new entries, ~15%–20% represent putative homologs with a conservative cutoff of $p < 10^{-9}$, thus identifying many conserved genes that are likely to share common functions with other well-studied organisms. Gene assemblies were also used to identify strain polymorphisms, examine stage-specific expression, and identify gene families. An interesting class of genes that are confined to members of this phylum and not shared by plants, animals, or fungi, was identified. These genes likely mediate the novel biological features of members of the Apicomplexa and hence offer great potential for biological investigation and as possible therapeutic targets.

[The sequence data from this study have been submitted to dbEST division of GenBank under accession nos.: *Toxoplasma gondii*: BG657138–BG661027, BI921045–BI921090, BI946571–BI946588, BM003839–BM004582, BM039066–BM040645, BM131233–BM133172, BM174962–BM176879, BM188953–BM189923, BM271559–BM271694. *Plasmodium falciparum*: BI670521–BI670830, BI813842–BI816393, BI936022–BI936312, BM273300–BM276553. *Sarcocystis neurona*: BE574328, BE574347, BE574384, BE574386, BE574409, BE574465, BE574508, BE574543, BE574561, BE574633, BE574689, BE574694, BE574723, BE635418–BE636244, BE574288–BE574724, BF323572–BF324064, BM252128–BM253024, BM303125–BM305293. *Eimeria tenella*: AI755306–AI758088, AI759179–AI759181, AI759254–AI759304, AI759182–AI759253, AI759305–AI759387, AI759463–AI759546, AI759388–AI759462, AI759547–AI759621, BE027133–BE028807, BF023640–BF023711, BF023609–BF023639, BG235514–BG235880, BG413067–BG413336, BG466192–BG467045, BG515959–BG517044, BG560819–BG562379, BG724474–BG725148, BI895002–BI896127, BM305294–BM306971, BM321464–BM322026. *Neospora caninum*: BF248514–BF249435, BF716421–BF717094, BF823742, BF823805–BF823813, BF823743–BF824633, BG235070–BG235513.]

The generation of expressed sequence tags (ESTs) provides a rapid means of gene discovery from single-pass sequencing of randomly selected cDNAs. This approach has been particularly useful for complex, model genomes including human (Hillier et al. 1996), rat (Scheetz et al. 2001), mouse (Marra et al. 1999b), fish (Clark et al. 2001), and rice (Ewing et al. 1999). One of the primary advantages of ESTs is that the identification of putative genes by BLAST comparisons (Altschul et al. 1990) enables researchers to begin biological analyses prior to the completion, or even initiation, of a full genome sequence. EST sequencing is likely to make its greatest impact on understudied genomes where little prior sequence data exists and

[9]**Corresponding author.**
E-MAIL sibley@borcim.wustl.edu; FAX (314) 362-3203.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.693203.

where full genome sequencing projects may not be undertaken in the near future. Parasites provide such a group, and previous EST projects have revealed the tremendous utility of this approach for gene discovery (Reddy et al. 1993; Chakrabarti et al. 1994; Wan et al. 1995; Ajioka et al. 1998; Manger et al. 1998b; Howe 2001). EST sequencing allows not only the rapid identification of abundantly expressed genes, it also provides data sets for informing phylogenetic analyses, examining strain diversity, and exploring developmentally regulated genes. Tools for recognizing and combining ESTs generated from the same gene into nonredundant assemblies in silico have recently been refined, improving the chances for establishing gene identities. Such identities, although only putative, enable rapid analysis of gene function, thus greatly facilitating traditional research approaches.

To further the process of gene discovery in protozoan parasites, we have undertaken large-scale EST sequencing projects for several apicomplexan parasites. The Apicomplexa is an ancient phylum of ~5000 species, all of which are parasitic (Levine 1970). Apicomplexans are most closely related to dinoflagellates and ciliates as shown by phylogenetic reconstructions based on small subunit ribosomal RNA sequences (Gajadhar et al. 1991; Escalante and Ayala 1994), and more recently, by examining conserved protein sequences (Baldauf et al. 2000). The age of the Apicomplexa predicts that many of their features will have diverged since their last common ancestry with the major eukaryotic kingdoms of plants, animals, and fungi. The relationships of major taxa within the Apicomplexa are depicted in Figure 1. Included are all major groupings in which EST or genome projects are presently underway. Additionally, the outgroups of ciliates (Paramecium, Oxytricha) and dinoflagellates (Prorocentrum, Symbiodinium) are shown for comparison. Apicomplexan parasites infect a wide range of vertebrate hosts and cause diseases of medical importance in humans, or veterinary importance in a range of domestic animals. In the present study, we have chosen the following organisms for study: *Plasmodium falciparum* and *Toxoplasma gondii*, which are both agents of human disease, and *Eimeria tenella*, *Neospora caninum*, and *Sarcocystis neurona*, which cause important diseases in agricultural and companion animals (Dubey 1977; Long 1993; Dubey and Lindsay 1996).

One of the major drawbacks of EST sequencing is the large number of database entries that are submitted separately to dbEST without extensive annotation. This complicates the problem of establishing which ESTs belong to a given gene, and whether similar ESTs belong to the same or closely related genes. Therefore, in addition to generating new ESTs from these organisms, we have clustered and assembled the resulting sequences into RNA consensus sequences and created a gene database that provides a variety of features for comparative analyses. Herein, we describe the creation of this database, illustrate several of its important features, and highlight several major features of gene content and expression in the Apicomplexa.

RESULTS AND DISCUSSION

General Strategy for Rapid Generation of ESTs

cDNA libraries were initially screened to ensure that they contained a high percentage of inserts (≥95% of ≥500 bp in size) and a reasonable diversity of sequences based on analyses of 50–100 clones per library. Sequences were subsequently obtained by single 5′ reads that were generated from randomly chosen cDNA clones. The numbers of ESTs generated for the organisms studied here are shown in Table 1. Sequences were analyzed for high quality, trimmed to remove vector sequences, and submitted to the dbEST division of GenBank, providing they met the criteria outlined in Methods. In general, the success rate, defined as the percentage of sequences that passed the submission criteria, ranged from 70%–80%, depending on the library in question. This strategy is rapid, economical, and amenable to high-throughput production. Sequencing of only the 5′ ends of cDNAs has been used previously in the *T. gondii* EST project (Wan et al. 1995; Ajioka et al. 1998; Manger et al. 1998b), and proved valuable for detecting gene similarities. In part, this is because of the early truncation of many cDNAs, combined with the relatively short 5′- UTRs, such that the regions sequenced tend to lie within the open reading frame, thus facilitating gene identi-

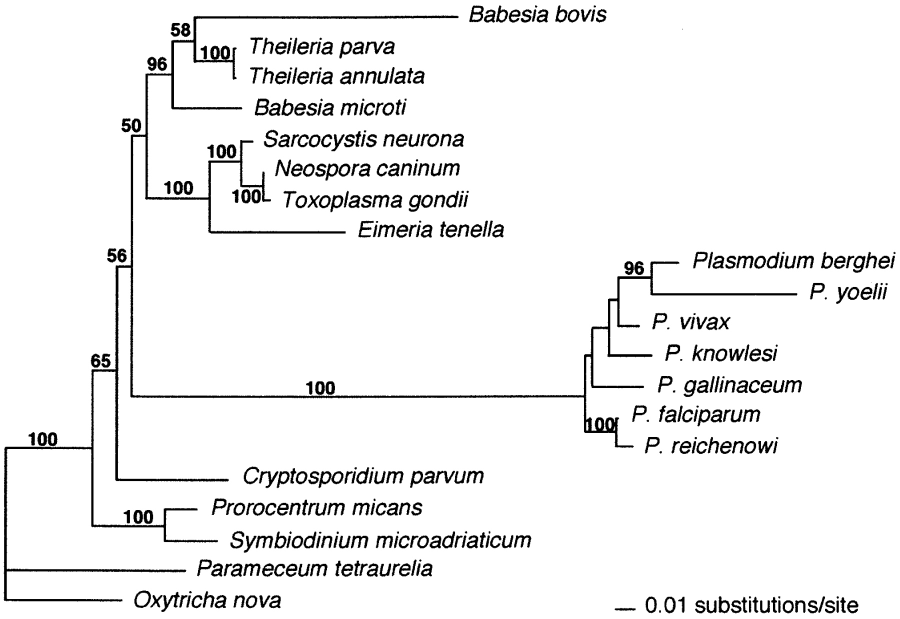


Figure 1 Unrooted distance phylogram generated from a neighbor-joining analysis of small subunit ribosomal genes. Organisms were chosen to illustrate relationships among the Alveolata with an emphasis on the Apicomplexa. The organisms chosen for study here include the three major branches of the Apicomplexa: *Plasmodium falciparum* representing the Haemosporidia; *Toxoplasma gondii, Neospora caninum, Sarcocystis neurona* representing the tissue-dwelling coccidia; *Eimeria tenella* representing the enteric coccidia; and *Babesia microti, Babesia bovis, Theileria parva,* and *Theileria annulata* representing the Piroplasms; *Symbiodinium microadriaticum* and *Prorocentrum micans* representing the Dinloflagellata; and *Parameceum tetraurelia* and *Oxytricha nova* representing the ciliates. *O. nova* was designated as the outgroup. Bootstrap percentages ≥50% are noted above the branches. The scale is as indicated.

Table 1. Distribution of Knowns and Unknowns for EST Assemblies

Organism	Number of ESTs[a]	Number of assemblies	Number of known genes[b] (%)	Number of putative genes[c] (%)	Number matching ProDom[d] (%)	Number matching CDD[e] (%)	Number of known–unknowns[f] (%)	Number unique[g] (%)
Taxoplasma gondii	23420	10585	72 (0.68)	1695 (16.01)	1633 (15.43)	1075 (10.16)	116 (1.10)	7829 (73.92)
Neospora caninum	3121	1388	2 (0.1)	223 (16.1)	238 (17.1)	167 (12.0)	9 (0.6)	1022 (73.63)
Sarcocystis neurona	4949	1445	0	219 (15.2)	215 (14.9)	171 (11.8)	16 (1.1)	1091 (75.50)
Eimeria tenella	13679	3425	8 (0.2)	592 (17.3)	529 (15.4)	453 (13.2)	54 (1.6)	2272 (66.33)
Plasmodium falciparum	10023	5800	243 (4.19)	886 (15.3)	1339 (23.09)	738 (12.7)	271 (4.67)	2989 (51.53)

The same number of significant digits were kept in the percentages shown for each entry.
[a] The number of ESTs for each organism reflects the content of dbEST/NCBI as of March 2002.
[b] 98% identity to a protein from the same species in SwissProt/PIR.
[c] $p < 10^{-9}$ similarity to a protein in SwissProt/PIR but excluding those assemblies in "known" and "known unknowns."
[d] $p < 10^{-9}$ similar to ProDom.
[e] $p < 10^{-9}$ similar to the Conserved Domain Database (CDD).
[f] $p < 10^{-9}$ similarity to a protein in SwissProt/PIR, but the best hit has description "hypothetical protein" or "unknown protein."
[g] No similarity found with $p < 10^{-5}$.

fication. A similar strategy was used for all of the projects described here. All of the EST sequences were initially compared with the mRNA database, and similarities were annotated at the time of submission.

Gene Assemblies and Generation of a Comparative Database

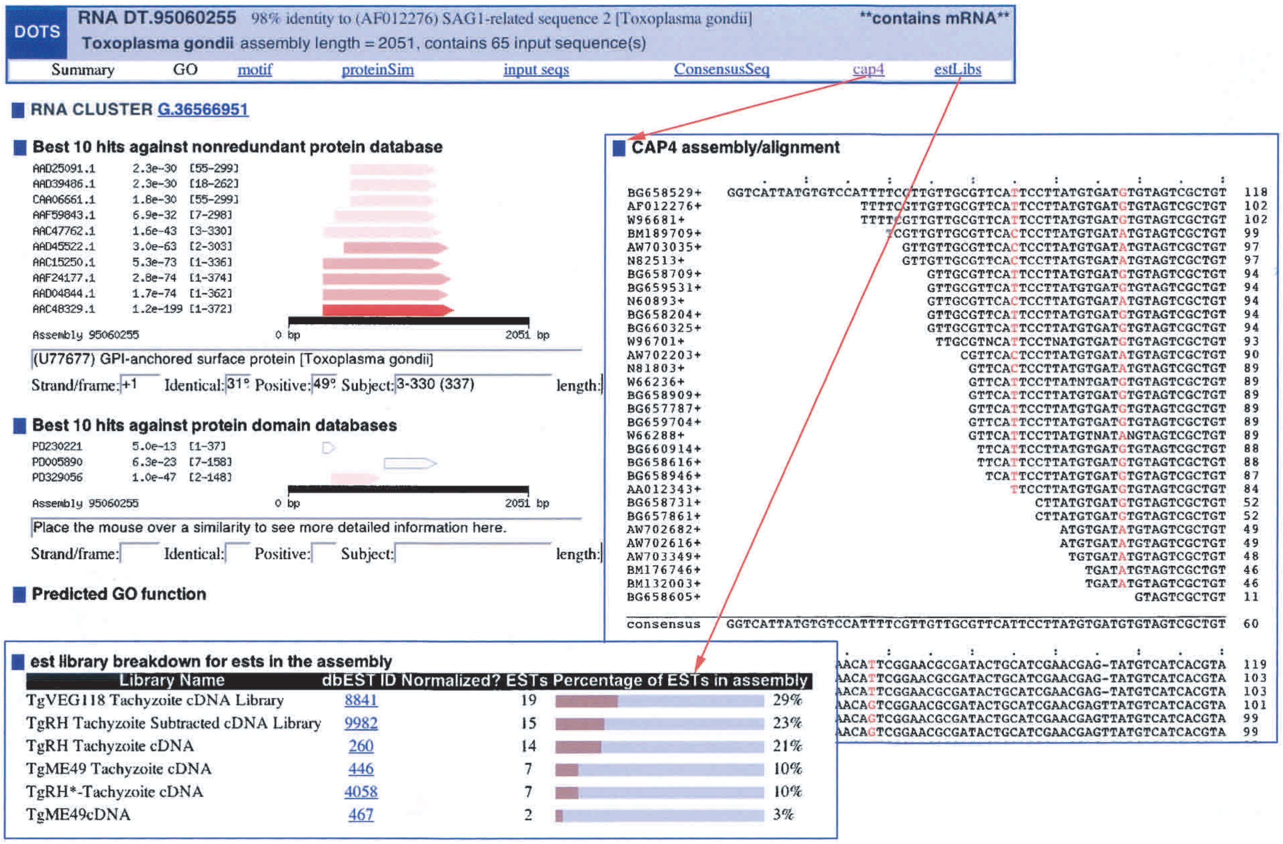
There are inherent difficulties in using raw EST sequences to identify genes because they are incomplete, highly redundant, and error prone. To alleviate these problems, we clustered the ESTs separately for each of the organisms shown in Table 1 to form consensus sequences. These clustered sequences were further grouped with their corresponding mRNAs in GenBank to generate "assemblies," which represent consensus sequences for a given transcript within each organism. To better facilitate comparative analyses, we used a relational database called GUS (Genomics Unified Schema; Davidson et al. 2001). This apicomplexan EST database comprises the separate assemblies from each species along with their annotations and information on the EST sources. The vast majority of these sequences were newly generated in the present project, although significant numbers of *T. gondii* ESTs have been reported on previously (Wan et al. 1995; Ajioka et al. 1998; Manger et al. 1998b), or in the case of *P. falciparum* were deposited into dbEST by others (Reddy et al. 1993; Chakrabarti et al. 1994). To make these data fully accessible and allow other researchers to ask their own questions, we have made the data and results available through a user-friendly Web interface called ApiESTDB (http://www.cbil.upenn.edu/paradbs-servlet/). The Web site was generated by Java Servlets and Perl programs previously developed for the AllGenes database (http://www.allgenes.org). ApiESTDB can be queried by information on the constituent ESTs (e.g., library, accession), key word searches on similarities to known genes and protein domains, and BLAST searches of a user-defined sequence. The database also supports Boolean queries that can be built by users on top of the available queries and is able to retain the history of a series of searches performed, whose results can be selectively combined by users. For each selected assembly, a summary page shows graphical displays of similar protein sequences and protein domains from BLAST results as well as links to alignment and library breakdown of the constituent ESTs (Fig. 2).

Identification of Genes

Once compiled, the assemblies for each organism were compared separately to SWISS-PROT/PIR using BLASTX similarity searches to identify previously known genes using a cutoff of 98% identity (referred to as "known genes" in Table 1). These genes represent previously characterized, full-length cDNAs from each of the organisms. Notably, the numbers of previously known genes are extremely low for *E. tenella* (8), *N. caninum* (2), and *S. neurona* (0), reflecting the limited extent to which these organisms have been studied. Although more genes have been previously described for *T. gondii* (72) and *P. falciparum* (243), these still represent a small fraction of the expected number of total genes.

To find putative homologs corresponding to conserved genes that have not been characterized in these parasites previously, the assemblies were compared with the SWISS-PROT/PIR (excluding hypothetical or unknown proteins) using BLASTX with a cutoff of p-value < 10^{-9}. A large number of statistically significant similarities were detected for each of the species, accounting for ~15%–17% of all assemblies (referred to as "putative" in Table 1). Similarly, comparisons to the protein domain families revealed ~15% of assemblies have similarities to ProDom (available at http://prodes.toulouse.inra.fr/prodom/doc/prodom.html), whereas ~10%–12% of the assemblies have significant similarities to Conserved Domain Database entries (CDD; available at http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) with a conservative cutoff of p-value < 10^{-9} (Table 1). A relatively small number of assemblies were most similar to GenBank nonredundant protein database (gbnr) entries annotated as "unknown", whereas the majority of genes in all the organisms examined were not similar to any previous gbnr entries and are classified as "unique" (Table 1). Some of these unique entries may represent 5′-UTRs rather than actual coding regions; however, previous analyses indicate that only a minority (≤10%) of ESTs actually extend into the 5′-UTR region in *T. gondii* genes (Ajioka et al. 1998; Manger et al. 1998b). The number of known genes is higher, and correspondingly the number of unique genes is lower in *P. falciparum*, which likely reflects the fact that the genome sequence is now complete and it has been extensively annotated (see http://PlasmoDB.org/ and links therein).

The search parameters used here for gene comparisons



DOTS **RNA DT.95060255** 98% identity to (AF012276) SAG1-related sequence 2 [Toxoplasma gondii] **contains mRNA**

Toxoplasma gondii assembly length = 2051, contains 65 input sequence(s)

RNA CLUSTER G.36566951

Best 10 hits against nonredundant protein database

```
AAD25091.1      2.3e-30  [55-299]
AAD39486.1      2.3e-30  [18-262]
CAA06661.1      1.8e-30  [55-299]
AAF59843.1      6.9e-32  [7-298]
AAC47762.1      1.6e-43  [3-330]
AAD45522.1      3.0e-63  [2-303]
AAC15250.1      5.3e-73  [1-336]
AAF24177.1      2.8e-74  [1-374]
AAD04844.1      1.7e-74  [1-362]
AAC48329.1      1.2e-199 [1-372]

Assembly 95060255                 0 bp                2051 bp
```

(U77677) GPI-anchored surface protein [Toxoplasma gondii]

Strand/frame: +1 Identical: 31° Positive: 49° Subject: 3-330 (337) length:

Best 10 hits against protein domain databases

```
PD230221        5.0e-13  [1-37]
PD005890        6.3e-23  [7-158]
PD329056        1.0e-47  [2-148]

Assembly 95060255                 0 bp                2051 bp
```

Place the mouse over a similarity to see more detailed information here.

Strand/frame: Identical: Positive: Subject: length:

Predicted GO function

CAP4 assembly/alignment

```
                 .    :    .    :    .    :    .    :    .    :    .    :
BG658529+    GGTCATTATGTGTCCATTTTCGTTGTTGCGTTCATTCCTTATGTGATGTGTAGTCGCTGT  118
AF012276+                    TTTTCGTTGTTGCGTTCATTCCTTATGTGATGTGTAGTCGCTGT  102
W96681+                      TTTTCGTTGTTGCGTTCATTCCTTATGTGATGTGTAGTCGCTGT  102
BM189709+                       TCGTTGTTGCGTTCACTCCTTATGTGATATGTAGTCGCTGT  99
AW703035+                         GTTGTTGCGTTCACTCCTTATGTGATATGTAGTCGCTGT  97
N82513+                           GTTGTTGCGTTCACTCCTTATGTGATATGTAGTCGCTGT  97
BG658709+                            GTTGCGTTCATTCCTTATGTGATGTGTAGTCGCTGT  94
BG659531+                            GTTGCGTTCATTCCTTATGTGATGTGTAGTCGCTGT  94
N60893+                              GTTGCGTTCACTCCTTATGTGATATGTAGTCGCTGT  94
BG658204+                            GTTGCGTTCATTCCTTATGTGATGTGTAGTCGCTGT  94
BG660325+                            GTTGCGTTCATTCCTTATGTGATGTGTAGTCGCTGT  94
W96701+                               TTGCGTNCATTCCTNATGTGATGTGTAGTCGCTGT  93
AW702203+                                CGTTCACTCCTTATGTGATATGTAGTCGCTGT  90
N81803+                                   GTTCACTCCTTATGTGATATGTAGTCGCTGT  89
W66236+                                   GTTCATTCCTTATNTGATGTGTAGTCGCTGT  89
BG658909+                                 GTTCATTCCTTATGTGATGTGTAGTCGCTGT  89
BG657787+                                 GTTCATTCCTTATGTGATGTGTAGTCGCTGT  89
BG659704+                                 GTTCATTCCTTATGTGATGTGTAGTCGCTGT  89
W66288+                                   GTTCATTCCTTATGTNATANGTAGTCGCTGT  89
BG660914+                                  TTCATTCCTTATGTGATGTGTAGTCGCTGT  88
BG658616+                                  TTCATTCCTTATGTGATGTGTAGTCGCTGT  88
BG658946+                                   TCATTCCTTATGTGATGTGTAGTCGCTGT  87
AA012343+                                      TTCCTTATGTGATGTGTAGTCGCTGT  84
BG658731+                                         CTTATGTGATGTGTAGTCGCTGT  52
BG657861+                                         CTTATGTGATGTGTAGTCGCTGT  52
AW702682+                                            ATGTGATATGTAGTCGCTGT  49
AW702616+                                            ATGTGATATGTAGTCGCTGT  49
AW703349+                                             TGTGATATGTAGTCGCTGT  48
BM176746+                                               TGATATGTAGTCGCTGT  46
BM132003+                                               TGATATGTAGTCGCTGT  46
BG658605+                                                     GTAGTCGCTGT  11
             ____________________________________________________________
consensus    GGTCATTATGTGTCCATTTTCGTTGTTGCGTTCATTCCTTATGTGATGTGTAGTCGCTGT  60

                 .    :    .    :    .    :    .    :    .    :    .    :
             AACATTCGGAACGCGATACTGCATCGAACGAG-TATGTCATCACGTA  119
             AACATTCGGAACGCGATACTGCATCGAACGAG-TATGTCATCACGTA  103
             AACATTCGGAACGCGATACTGCATCGAACGAG-TATGTCATCACGTA  103
             AACAGTCGGAACGCGATACTGCATCGAACGAGTTATGTCATCACGTA  101
             AACAGTCGGAACGCGATACTGCATCGAACGAGTTATGTCATCACGTA  99
             AACAGTCGGAACGCGATACTGCATCGAACGAGTTATGTCATCACGTA  99
```

est library breakdown for ests in the assembly

Library Name	dbEST ID	Normalized?	ESTs	Percentage of ESTs in assembly
TgVEG118 Tachyzoite cDNA Library	8841		19	29%
TgRH Tachyzoite Subtracted cDNA Library	9982		15	23%
TgRH Tachyzoite cDNA	260		14	21%
TgME49 Tachyzoite cDNA	446		7	10%
TgRH*-Tachyzoite cDNA	4058		7	10%
TgME49cDNA	467		2	3%

Figure 2 The summary page for assembly DT.95060255. For each assembly, there is a summary page with links in the heading to detailed information about this assembly. The top 10 BLAST similarities to protein and domain databases are shown graphically on the summary page, whereas "proteinSim" and "motif" are links to their tabular presentation, respectively. "input seqs" links to constituent ESTs of this assembly; "ConsensusSeq" links to the consensus sequence of this assembly; "cap4" links to the alignment of ESTs from the cap4 assembling program with SNPs highlighted; and "estLibs" links to the library breakdown of ESTs in this assembly.

were quite selective, and it is likely that more sensitive searches would identify additional similarities. ApiESTDB allows investigators to perform a variety of self-directed BLAST searches and the standard parameters (*E*, *S*, matrix) can be set by the user. Consequently, the analysis performed here should be considered as a relatively selective assessment of gene similarities rather than an exhaustive analysis.

Redundancy

One inevitable consequence of the nonuniform abundance of mRNAs for different genes is that some genes are overrepresented, whereas others are relatively rare in cDNA libraries. This bias can be further enhanced by amplification or differences in cloning efficiency during construction of cDNA libraries. To analyze the extent of redundancy in cDNA libraries of each of the organism studied here, we compared assemblies generated from nonsubtracted libraries from each organism. Assemblies were analyzed by determining the number of ESTs they contained as a percentage of all ESTs for that organism after normalization for the total sample sizes. When presented graphically, it is evident that the inherent redundancy in *P. falciparum* and *T. gondii* was relatively low (Fig. 3), where >90% of ESTs occur in assemblies that each constitute <0.1% of all ESTs. In contrast, the redundancy of *N. caninum*, *E. tenella*, and *S. neurona* libraries was considerably higher (Fig. 3) and up to 20% of assemblies each contain >1% of all ESTs. This difference may in part be a reflection of real biological differences in mRNA abundances, but is also likely due in part to differences in amplification of abundant transcripts during library construction. The procedures used here for library generation from different organisms were highly similar. For example, all libraries used size-selected cDNAs to avoid problems with overrepresentation of small transcripts. Nonetheless, some of the observed differences in transcript abundance may have resulted from artifacts in library construction.

The redundancy inherent in cDNA libraries is beneficial in the generation of assemblies as overlapping ESTs from a single gene can be aligned and compiled to generate a consensus sequence in silico. Comparison of multiple ESTs from a given assembly is useful for identifying the boundaries of open reading frames, alternative splicing, and strain-specific polymorphisms. Putative alternatively spliced transcripts were identified by clustering the consensus sequences of assemblies with a minimum cutoff of 95% identity over >150 bp length and cross-linking related assemblies to each other. In the case of *T. gondii*, libraries were chosen to represent each of the three predominant genetic lineages (Howe and Sibley 1995), and single-nucleotide polymorphisms are highlighted in the alignments for a given assembly (see Fig. 2). Additionally, the source library for each EST is retained, allowing easy determination of the relative abundance in a given life cycle stage or strain type.

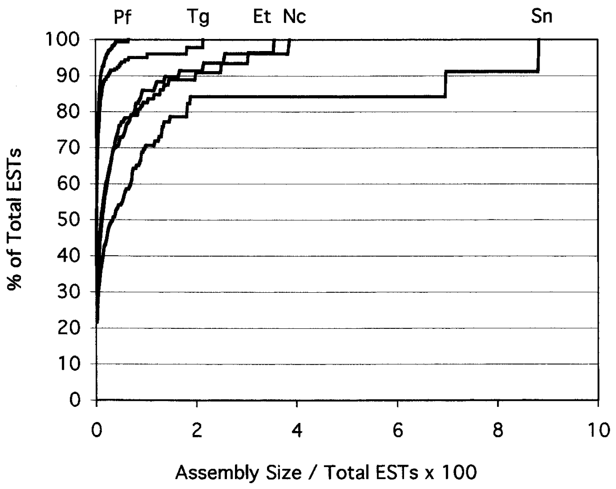


Figure 3 Redundancy of apicomplexan ESTs. The percentage of the total ESTs (plotted on the *Y*-axis) is plotted as a function of increasing assembly sizes (expressed as a percentage of total ESTs on the *X*-axis). Significantly greater redundancy was observed in *Neospora caninum* (Nc), *Eimeria tenella* (Et), and *Sarcocystis neurona* (Sn) relative to *Toxoplasma gondii* (Tg) and *Plasmodium falciparum* (Pf). The normalized cluster size of EST assemblies was determined by comparing the number of ESTs within each assembly to the total number of ESTs for that organism.

Abundantly Expressed Genes

Another potential use of redundancy is to identify highly expressed genes that may represent particularly important biological pathways for the organism. To identify such overexpressed genes, we analyzed the top 25 assemblies ranked by the number of ESTs contained in nonsubtracted libraries for each of the organisms and characterized the top hits based on the classification of functional categories as defined by NCBI as part of the conserved orthologous genes (COGs) analysis (available at http://www.ncbi.nlm.nih.gov/cgi-bin/COG/). In addition to the standard categories, we added several categories to reflect the abundance of genes known only from the apicomplexans, such as those encoding secretory proteins as well as surface antigens. Several distinct patterns emerged from this analysis, which are summarized in Figure 4. A listing of these assemblies can be found at http://www.cbil.upenn.edu/paradbs-servlet/pub/SupTable1Top25.htm. First, it was evident that *P. falciparum* was by far the most diverse with the largest number of categories being represented by the group of 25 most abundant EST assemblies. Second, the patterns for *T. gondii* and *N. caninum* were quite similar and were dominated by secretory proteins including those found in rhoptries, dense granules, and micronenes. This likely is in part a reflection of their close evolutionary similarity (Fig. 1), but is also a reflection of the biological specialization of the Apicomplexa, as such secretory components were also abundant in *E. tenella* and *P. falciparum*. These secretory components are sequentially discharged during cell invasion and participate in attachment, entry, and intracellular survival of the parasites (Dubremetz et al. 1993; Carruthers and Sibley 1997). A second major class of abundant ESTs was a surface antigen family known as SAGs, first identified in *T. gondii* (Manger et al. 1998a; Boothroyd et al. 1997; Lekutis et al. 2000). Members of this family were identified in *N. caninum* as well as in *S. neurona*, and a related, but distinct family of cysteine-rich surface antigens was abundant in *E. tenella*. Finally, the pattern of highly expressed ESTs in *S. neurona* was highly unusual in that all but three of the top 25 EST assemblies were classified as unique: of the remaining three, two are similar to SAGs and the third is similar to the ubiquitin-like protein SMT3.

Stage-Specific Gene Expression

Apicomplexan life cycles involve a number of developmental adaptations that are designed to accommodate changes in the environment. These include changes between different hosts, either vertebrate or invertebrate, and between the host and extracellular environments (i.e., oocyst shedding). We analyzed the number of genes that are expressed in a stage-specific manner based on an arbitrary cutoff as well as statistical significance of their differential expression levels as reflected by abundance of ESTs at each stage (Table 2). A listing of these stage-specific assemblies can be found at: http://www.cbil.upenn.edu/paradbs-servlet/pub/SupTable2Stage.htm. In the case of *P. falciparum*, we obtained ESTs only from the asexual forms that replicate in blood cells and from gametocytes and did not include mosquito or liver stages. More than 80% of assemblies were stage-specific as judged by being five-

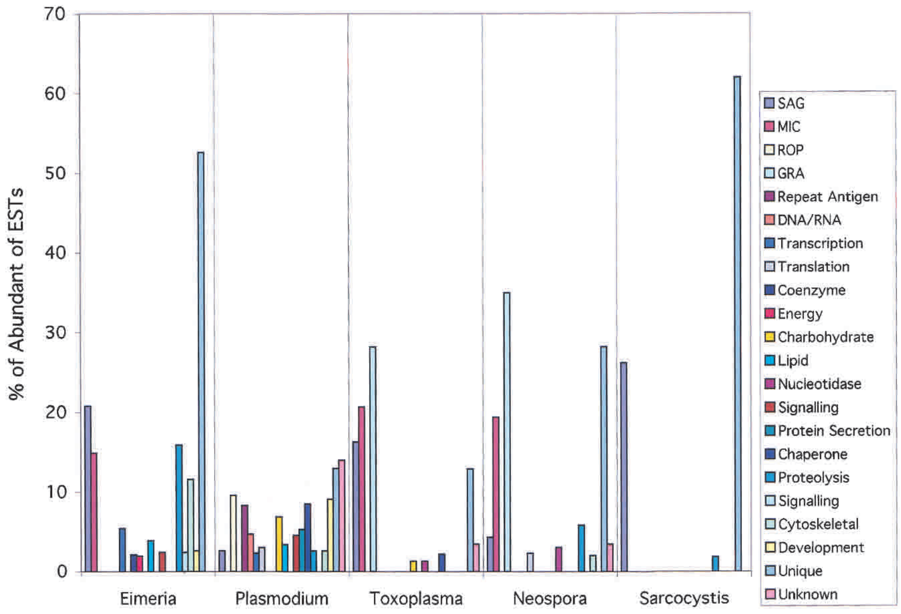


Figure 4 Most abundant 25 assemblies for each organism. Data are plotted as number of all ESTs contained within a particular category for each of the top 25 assemblies. Functional groupings were assigned using the categories established for the COGs database with the addition of categories for surface antigens (SAG), and secretory proteins from the micronemes (MIC), rhoptries (ROP), and dense granules (GRA).

Table 2. Stage-Specific Expression of Gene Families in Apicomplexans

Organism	Tachyzoite/merozoite	Oocyst/sporozoite	Gametocyte	Bradyzoite	Constitutive
Taxoplasma gondii					
Fivefold	234 (41.2)[a]	66 (11.6)	ND	69 (12.1)	199 (35.0)
$P \leq 0.05$	13 (2.3)	48 (8.5)		32 (5.6)	475 (83.6)
Plasmodium falciparum					
Fivefold	60 (29.6)	ND	108 (53.2)	NA	35 (17.2)
$P \leq 0.05$	14 (6.9)		10 (4.9)		179 (88.2)
Eimeria tenella					
Fivefold	65 (46.4)	51 (36.4)	ND	NA	24 (17.1)
$P \leq 0.05$	15 (10.7)	16 (11.4)			109 (77.9)

Based on assemblies where $N \geq 5$ ESTs from nonsubtracted libraries. For *E. tenella,* there were 140 assemblies included; for *P. falciparum,* there were 203 assemblies included; for *T. gondii,* there were 568 assemblies included. Two different cutoffs were used, fivefold difference or $P \leq 0.05$. NA, not applicable; ND, not done.

[a]Values expressed as total asssemblies (% of total).

fold different; however, only 11% of these were statistically significant ($p \leq 0.05$). In the case of *E. tenella*, the two predominate developmental stages are intracellular merozoites, that replicate within the intestine, and oocysts, which survive in the environment and contain sporozoites that are responsible for transmission. In comparing these two different stages, again >80% of assemblies were stage-specific by the fivefold criteria, whereas ~20% of these were statistically significant ($p \leq 0.05$). In the case of *T. gondii*, we were able to generate cDNA libraries from two of the major replicating forms, tachyzoites, bradyzoites, and from oocysts, the extracellular form responsible for environmental transmission. Approximately 2.3%, 5.6%, and 8.5% assemblies were specific to each of these stages, respectively, indicating that over all ~16% of genes are expressed in a stage-specific manner based on the $p \leq 0.05$ cutoff (Table 2). Little is known about the molecular mechanisms of stage-specific expression. However, these life cycle stages must survive in very different environments, and presumably these changes in gene expression underlie important biological adaptations that are specifically needed in each of these niches. One caveat to the analyses presented here is that the relative abundance of a given transcript in a particular library may have been influenced by procedures in library construction. More sensitive methods for examining stage-specific expression, such as microarrays (Lashkari et al. 1997; Eisen and Brown 1999; Cummings and Relman 2000) or SAGE (Velculescu et al. 1995), will be needed to validate and extend these findings to other gene families. The availability of an EST gene database will facilitate such analysis by providing target genes for microarray studies, and by providing sequences for comparison to SAGE tags.

Phylogenetic Comparison of Gene Homologies

In classifying the orthologs identified by EST sequencing, it was of interest to determine the phylogenetic associations of the closest neighbor. Such information could be useful for establishing gene origins and for identifying genes with restricted phylogenetic distributions versus those that are widely dispersed. Thus, we further classified those assemblies that have homologs obtained from Table1 based on the phylogenetic origin of their putative orthologs (see http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/; Table 3). Surprisingly, only 1%–2% of all assemblies have best non-self (not from the same organism) similarity to another apicomplexan. In part, this likely reflects the relatively small coverage of most of these genomes and underrepresentation of proteins from these organisms in the database. Approximately 3%–5% showed the highest similarity outside the phylum Apicomplexa to a plant gene, whereas 5%–10% were most similar to animals (Table 3). Previous studies have also highlighted the fact that many metabolic enzymes in *T. gondii* are plant-like (Dzierszinski et al. 2001). These plant-like genes may have originated from an algal endosymbiont that is still represented by the remnant apicoplast (Kohler et al. 1997; Roos et al. 1999). Similarities to archaea and fungi were less common, and these may represent important examples of convergence or of lateral gene transfer. One of the major im-

TABLE 3. Closest Phylogenetic Ortholog for Apicomplexan Gene Assemblies

Organism	Best nonself match to Api	Best non-Api match to plant	Best non-Api match to fungi	Best non-Api match to bacteria or archeae	Best non-Api match to metazoan
Taxoplasma gondii (10,597)[a]	224[b] (2.11)	425 (4.01)	212 (1.99)	126 (1.19)	703 (6.63)
Neospora caninum (1394)	64 (4.6)	65 (4.7)	23 (1.6)	14 (1.0)	81 (5.8)
Sarcocystis neurona (1445)	30 (2.1)	40 (2.8)	22 (1.5)	3 (0.2)	147 (10.2)
Eimeria tenella (3439)	92 (2.7)	164 (4.77)	100 (2.91)	51 (1.5)	264 (7.68)
Plasmodium falciparum (5992)	97 (1.6)	235 (3.92)	217 (3.62)	85 (1.4)	320 (5.34)

[a]Total number of assemblies including mRNA singletons.

[b]Number of assemblies in category (%).

plications of the presence of plant-like genes, and the less common, but nonetheless important, prokaryotic-like genes, is that they identify metabolic pathways in apicomplexans that may be significantly different from those of their mammalian hosts. Thus, these pathways represent prime candidates for development of new therapeutics that might be expected to offer a high degree of selectivity. It is possible to further analyze such candidates in terms of putative function and domain structure using ApiESTDB to design specific searches and queries.

Identification of Genes Restricted to the Apicomplexa

One of the primary reasons to undertake the present project was to identify genes that are restricted to the Apicomplexa, as these may mediate their unique biology. We performed two comparisons to identify apicomplexan-specific genes, dividing them into those that have a low significance match to SWISS-PROT/PIR and those that are completely unique to the Apicomplexa. For the first set, we compiled a list of assemblies that were similar to at least one other apicomplexan with a p-value < 10^{-9} based on comparison to SWISS-PROT/PIR and have no similarity outside the Apicomplexa with a p-value < 10^{-5}. Only assemblies with a putative match to SWISS-PROT/PIR of $p < 10^{-9}$, equal to ~20% of all assemblies, were included in this analysis. A total of 105 assemblies were found to be apicomplexan-specific, representing 62 genes of known or putative function. These are summarized in Figure 5, and a complete listing can be found at http://www.cbil.upenn.edu/paradbs-servlet/pub/SupTable3Api.htm. The assemblies fall into several broad categories including (1) those that are specifically restricted to the Apicomplexa; (2) those that are also known from other organisms, yet are significantly more similar within the Apicomplexa. Most apicomplexan-specific genes are related to their unique apical specialization and encode such components as secretory organelles, surface antigens, and developmental components. Included in this category are genes encoding secretory proteins such as apical membrane antigen 1 (AMA-1) that is found in *P. falciparum*, *T. gondii*, and *E. tenella*. AMA-1 is present on the parasite cell surface and has been implicated in host cell invasion by *P. falciparum* (Peterson et al. 1989; Waters et al. 1990) and *T. gondii* (Donahue et al. 2000; Hehl et al. 2000). Also included are other microneme proteins, likely involved in cell recognition (Soldati et al. 2001), the major surface antigens (SAGs; Boothroyd et al. 1997), and a number of proteins characterized from sporozoites or oocysts. Another highly conserved, yet unique apicomplexan protein is the small myosin known as TgMyoA in *T. gondii* (Hettman et al. 2000) and MyoA in *P. falciparum* (Pinder et al. 1998). This motor protein plays an important role in the unique form of motility of this group of parasites (Sibley et al. 1998). Additionally, a number of metabolic enzymes are found on this list including a unique apyrase known as nucleoside triphosphate hydrolase (NPTase), first characterized in *T. gondii* (Asai et al. 1995) and also found in *N. caninum* (Asai et al. 1998). In further sequencing, a putative homolog of the NTPase has been identified from *S. neurona* (D. Howe, unpubl.), indicating that this enzyme may be more widely contained in the tissue coccidians.

Several conserved enzymes and signaling molecules common to many taxa are also found on the apicomplexan-specific list, for example, hexokinase, 14-3-3 protein, and hypoxanthine phosphoribosyltransferase. Although not restricted to parasites, they are more highly conserved within the Apicomplexa, indicating that they fulfill specific roles that are unique to the members of this phylum. These similarities are not simply caused by these organisms being closer to each other phylogenetically, as many highly conserved genes found in this group do not fit the criteria for being apicomplexan-specific. As such, the apicomplexan-specific features within conserved genes may provide potential targets for development of selective inhibitors.

Because of the limited number of known genes in this phylum, the majority of the assemblies do not have significant similarity to the databases. The similarities among these assemblies from different members of this group, however, will provide hints on their importance in apicomplexan-specific physiology and prioritize them for functional characterization. Therefore, we carried out TBLASTX comparisons among all the apicomplexan assemblies and to human–mouse gene sequences. We identified those apicomplexan assemblies that were not similar to known proteins or human–mouse assemblies (p-value < 10^{-5}), but were similar to at least another apicomplexan assembly (p-value < 10^{-9}). This gave 326 assemblies from *T. gondii*, 173 from *N. caninum*, 28 from *S. neurona*, 36 from *P. falciparum*, and 82 from *E. tenella*. These assemblies represent unknown genes that are conserved among Apicomplexa but not found in model animal host, human or mouse. A list of these assemblies can be found at http://www.cbil.upenn.edu/paradbs-servlet/pub/SupTable4Api/.

The creation of a relational database to house gene assemblies

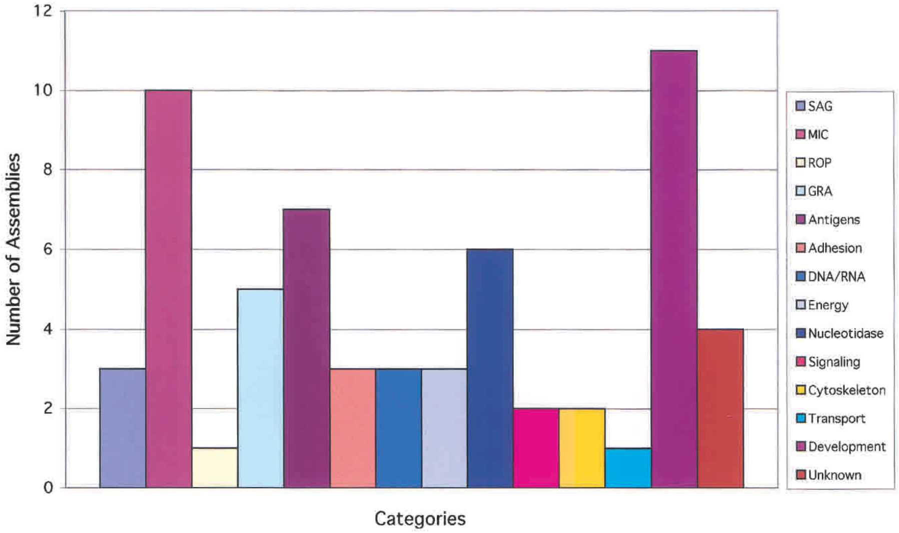


Figure 5 Apicomplexan specific gene families categorized by function. The majority of apicomplexan-specific genes are related to surface antigens, secretory proteins, and developmentally regulated genes. A total of 105 gene assemblies, representing 62 distinct genes, are included. Functional categories are the same as those listed in Figure 4.

built from EST sequences should greatly facilitate analysis of gene function, expression, and relatedness among the Apicomplexa. The architecture of ApiESTDB allows it to be periodically updated by inclusion of new data from ongoing sequencing projects in these and related organisms. ApiESTDB is also highly complementary to the recently completed genome sequences for several malaria species (Carlton et al. 2002; Gardner et al. 2002; http://www.nature.com/nature/malaria/index.html). Comparison of ESTs sequences with the full genome is provided by a separate resource called PlasmoDB (Kissinger et al. 2002; http://PlasmoDB.org/). One of the remaining deficiencies in our understanding of the Apicomplexa is that very few sequences are available for related deep-branch organisms such as ciliates, dinoflagellates, and even for some apicomplexan groups such as gregarines. An increased investment in sequencing from these taxa would be highly informative in terms of defining genes that are restricted to specific branches of the Apicomplexa, especially those known to be economically or medically important parasites of animals and man.

METHODS

Parasite Culture and Library Construction

T. gondii

cDNA libraries constructed in the type I RH strain or the type II ME49 strain (B7 clone previously referred to as PDS) have been described previously (Ajioka et al. 1998). Tachyzoites of the type III strain VEG, an isolate from an AIDS patient (obtained from Jack Remington), were grown in human foreskin fibroblast (HFF) cell monolayers as described previously (Roos et al. 1994). Tachyzoites were separated from host cells by passage through 3.0-micron membrane filters (Nucleopore, Inc.) and centrifugation (Howe and Sibley 1995). mRNAs were isolated from freshly isolated parasites treated with TRIZOL (Invitrogen, Inc.). cDNAs were synthesized by oligo(dT) priming from poly(A) mRNA, sized selected, and directionally cloned into a Uni-ZAP XR vector (Stratagene).

To obtain libraries from the oocyst stage of the life cycle, mRNA was isolated from partially sprorulated and fully sporulated oocysts that were produced in cats. Oocysts were purified by sucrose gradient flotation and chlorox treatment as described (Tilley et al. 1997). Poly(A) mRNAs were converted to cDNA using the template-switching PCR strategy (SMART cDNA, Clontech Inc.) Size-selected cDNAs, containing *Sfi*I linker ends, were ligated into a modified pBluescript vector containing directional *Sfi*I sites and electroporated into DH10B *E. coli* cells.

N. caninum

Tachyzoites of the Nc-1 isolate (Lindsay and Dubey 1989) were grown in HFF fibroblasts in DMEM containing 10% horse serum as described previously. Tachyzoites were purified as described above for *T. gondii*. cDNAs were synthesized by oligo(dT) priming from poly(A) mRNA, size-selected, and directionally cloned into a Uni-Zap XR vector (Stratagene).

S. neurona

Merozoites of the Sn3 strain were grown in bovine turbinate cells in DMEM containing 10% fetal bovine serum. Parasites were purified as described above for *T. gondii*. cDNAs were synthesized by oligo(dT) priming from poly(A) mRNA, size-selected, and directionally cloned into a Uni-Zap XR vector (Stratagene).

E. tenella

Oocysts of the LS18 strain were obtained by fecal flotation, and merozoites were obtained from ceacal scrapings from infected chickens. Parasites were separated from host-cell material by repeated centrifugation. cDNAs were synthesized by oligo(dT) priming from (polyA) mRNA, size-selected, and directionally cloned into a Lambda ZapII vector (Stratagene).

P. falciparum

A Clone 3D7 infected erythrocytes were used for isolation of poly(A) mRNA by acidic guanidium-phenol chloroform extraction and poly(AT)-tract mRNA isolation (Promega). cDNAs were constructed by oligo(dT) priming of poly-mRNA, size-selected, and directional cloning into a Lambda ZapII vector (Stratagene). This library has been deposited with the MR4 collection as entry MRA-299. Gametocyte-stage-enriched cDNAs were synthesized from poly(A) mRNA by oligo(dT) priming, size-selected, and directionally cloned into a Uni-Zap-XR vector (Statagene). This library has been deposited with the MR4 collection as entry MRA-101 (at http://www.niaid.nih.gov/dmid/malaria/malrep/). Both *P. falciparum* libraries were amplified as phage prior to being converted to plasmids for sequencing as described below.

Clone Propagation and DNA Isolation

Packaged phage were used to infect XL1-Blue cells, and phagemids were rescued by ExAssist helper phage and propagated as plasmids in SOLR *E. coli* cells (Stratagene). Because of problems with stability, some clones were later converted into DH10B Gene Hogs (Invitrogen Inc.) *E. coli* by electroporation of purified plasmids prepared from colonies grown on LB-agar plates containing 100 µg/mL ampicillin. DH10B cells were plated onto LB-agar plates containing 100 µg/mL ampicillin, and individual colonies were picked to 96-well plates. Cultures were grown overnight in Terrific Broth containing 8% glycerol and 100 µg/mL of ampicillin with shaking to an OD of >0.6. Plates were stored at −80°C until used for expansion and cycle sequencing.

Then 5 µL of each cDNA glycerol stock in a 96-well microtiter plate was transferred by a Tango (Robbons Scientific) robot to each well of a 96-well Beckman block, containing 1 mL of Terrific Broth (Difco) with the appropriate antibiotic. Blocks were incubated at 37°C for 24 h with agitation at 295 rpm. Cells were pelleted and processed as described previously (Marra et al. 1999a), using a high-throughput 96-well microwave protocol.

DNA Sequencing

DNA was sequenced using BigDye terminator chemistry (ABI). Sequencing reactions were run on either ABI 377 fluorescent sequencers or ABI 3700 capillary sequencers. DNA electrophoretograms (traces) were evaluated using the computer programs GelMinder and EST_OTTO. All data passing preliminary evaluation were sent to the EST Computer Analysis Group at Washington University.

Processing and Annotation

Raw sequence data from the ABI 3700 were automatically processed to (1) generate basecalls and per base quality values; (2) determine high-quality start and stop positions; (3) trim flanking vector sequences; (4) mask repetitive elements; and (5) remove vector, *E. coli*, rRNA, and/or mitochondrial contamination. The resulting high-quality, trimmed sequences were then annotated with similarity information and library details, and then submitted to dbEST/GenBank.

The methods have been previously described (Hillier et al. 1996) with the following modifications. Vector sequences were trimmed using the programs WEP (W. Gish, unpubl.) and BLASTN2 (W. Gish, unpubl.), in which *S* (score) = 133, *S*2 (minimum reported score) = 133, M (match) = 5, N (mismatch) = -11, W (word size) = 7, R (gap extension penalty) = 11, Q (gap initiation penalty) = 11, and $E2$ (minimal reported *e*-value) = 0.5. WEP also served to identify incorrect adaptor sequences. Repetitive elements were identified and masked using blastx_and_mask (G. Miklem, unpubl.), which uses BLASTX (S = 50) to compare with a database of *T. gondii* repeat elements (GenBank accession nos. M57916, M57917, M57918, M57919, X60240, X60241, X60242, and X75429) translated in all six frames. Because no specific repeat element database was available for *N. caninum*, *S. neurona*, or *E. tenella*, the *T. gondii* repeat database was used for screening. No sequences were found to match this repeat database at the significance level stated above. *P. falciparum* was not screened for repeat sequences. The programs TANDEM and INVERTED (R. Durbin, unpubl.) were used to mask local tandem and inverted repeats, and DUST (R. Tatusov and D. Lipman, unpubl.) was used to mask low entropy sequence in *T. gondii*, *E. tenella*, *N. caninum*, and *S. neurona*. *P. falciparum* was not subject to low entropy filters (because of its expected low GC content). Sequences determined to be vector (BLASTN2 S = 170, gap $S2$ = 150, M = 5, N = -11, Q = 11, R = 11, W = 10, $E2$ = 0.5 against a vector subset of GenBank), *E. coli* (BLASTN2 S = 133, $S2$ = 133, M = 5, N = -11, Q = 11, R = 11, W = 10, $E2$ = 0.5 against an *E. coli* subset of GenBank), structural RNA (BLASTN2 S = 170, gap $S2$ = 150, M = 5, N = -11, Q = 11, R = 11, W = 10, $E2$ = 0.5 against the gbrna subset of GenBank), or human or mouse mitochondrial (BLASTN2 S = 170, gap $S2$ = 150, M = 5, N = -11, Q = 11, R = 11, W = 10, $E2$ = 0.5 against GenBank hummtc and musmtc) were not submitted to the public databases or included in further analysis.

Because all of the parasites examined here were grown in the presence of host cells, it is possible that some ESTs actually correspond to host mRNAs. To determine the frequency of this we compared the ESTs to sequences in GenBank nr (restricted to the following taxa: *Homo*, *Mus*, *Xenopus*, *Gallus*, *Rattus*, *Bos*, *Danio*, *Sus*) using BLASTN and a cutoff of >95% identical over a region of >100 nt. The following percentages of sequences matched this criteria: *Toxoplasma* (1.0%), *Eimeria* (0.05%), *Neospora* (0.6%), *Sarcocystis* (1.3%), and *Plasmodium* (0.75). Because it can be difficult to determine true contaminants from highly similar but distinct genes, all ESTs were submitted to GenBank with only a general qualifying statement "that a small proportion of ESTs represent host contaminants and investigators should be aware of this possibility."

Identification of Gene Similarities Prior to Submission

Similarities to proteins were identified using BLASTX (S = 100, M = PAM120, V = 0, W = 4, T = 17) searches against SWIR (release 21), which is a nonredundant protein database containing sequences culled from PIR, SWISS-PROT, and a database of predicted *Caenorhabditis elegans* proteins called WORMPEP (E. Sonnhammer, unpubl.). These identities were annotated at the time of entry and appear in the field "putative IDs" on some of the NCBI entries. ESTs generated here are also listed at http://genome.wustl.edu/EST/. ESTs with a high level of similarity to a known gene but on the opposite strand were annotated as such during submission. This fraction accounted for the following percentages of ESTs: *Eimeria* (4.9%), *Neospora* (1.9%), *Sarcocystis* (1.3%), *Plasmodium* (7.5%), and *Toxoplasma* (0.75%).

Clustering and Assembly of EST/mRNA Sequences

EST/mRNA sequences from each organism of interest were first cleaned by detecting and removing vector sequences using cross_match (using the phrap package) and the GenBank vector database, removing tailing poly(A) and leading poly(T) sequences and removing poor quality ends where the percentage of Ns in a 20-bp window exceeded 20%. Sequences shorter than 50 bp following this process were marked as "low_quality" and were not considered in the later steps of clustering. Sequences were then clustered by running an "all-against-all" WU BLASTN (W. Gish, unpubl.; http://blast.wustl.edu) comparison with parameters N = -10, M = 5 to limit extension of matches into poor-quality regions. Clusters were formed by a connected components analysis of all the BLASTN matches with minimum cutoff values of 92% identity and 40-bp length and having two ends matching. The clusters were assembled to form consensus sequences using the CAP4 algorithm (at http://www.paracel.com/publications/cap4_092200.pdf). The CAP4 alignments were decomposed into constituent parts and stored in the GUS relational database housed at the Center for Bioinformatics, University of Pennsylvania (Davidson et al. 2001). Assemblies were reverse complemented if assembly orientation was inconsistent with mRNA orientation and EST clone end assignment.

Annotation of Consensus Sequences

Consensus sequences of the assemblies were annotated after comparing them to the GenBank nonredundant (gbnr) protein database (as of December 5, 2001) using WU-BLASTX (W. Gish, unpubl.; http://blast.wustl.edu) with the parameters of $-$wordmask = seg + xnu, W = 3, T = 1000, and similarities $p < 10^{-5}$ were loaded into the GUS database. Protein domain similarities were assigned for the consensus sequences by comparing them to the ProDom database (version 2001.1; at http://prodes.toulouse.inra.fr/prodom/doc/prodom.html) using BLASTX (W. Gish, unpubl.; http://blast.wustl.edu) with the parameters of $-$wordmask = seg + xnu, W = 3, T = 1000, and to the CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) database (as of June 27, 2001) using RPS-BLAST (NCBI-BLAST) with the parameters $-a$ 2, $-e$ 0.1, $-p$ F. Similarities better than $p < 10^{-5}$ were loaded into the GUS database. The protein domain similarities were used to predict GO functional categories for the consensus sequences as previously described (Schug et al. 2002).

Identification of Phylogenetically Restricted Genes

Consensus sequences of the apicomplexan assemblies were compared with the gbnr protein database (as of December 5, 2001) using BLASTX (W. Gish, unpubl.; http://blast.wustl.edu). NCBI taxonomy was used to determine the phylogeny of protein sequences (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/). Assemblies that did not have similarity to non-apicomplexan proteins with $p < 10^{-5}$ or better, but which did have similarity $p < 10^{-9}$ to another apicomplexan entry from SWISS-PROT/PIR were identified as apicomplexan-specific. The best protein similarity was used to annotate their putative function. Consensus sequences from the apicomplexan organisms were also compared with each other by an all-against-all TBLASTX analysis, and to human/mouse gene sequences obtained from the Allgenes database (http://www.cbil.upenn.edu/downloads/DoTS/AllGenesRelease4.0/) by TBLASTX (W. Gish, unpubl.; http://blast.wustl.edu). The parameters used were $-$wordmask = seg + xnu, W = 3, T = 1000. Assemblies that were similar to at least one other assembly from an apicomplexan organism with $p < 10^{-9}$ but were not similar to human–mouse genes with $p < 10^{-5}$, were identified as putative apicomplexan-specific genes with unknown function.

Redundancy

ESTs from nonsubtracted libraries from each organism of interest were used to analyze redundancy and identify the most abundant genes. Because the total number of ESTs and assemblies varied widely, it was first necessary to normalize the data to allow direct comparisons between organisms. The sizes of the assemblies (number of ESTs from nonsubtracted libraries) from each organism were normalized as a percentage of the total number of ESTs included for that organism. Assemblies were then classified based on their normalized content size as a percentage of all ESTs for that organism. Redundancy was estimated by plotting the percentage of all total EST (*Y*-axis) versus the normalized assembly size as a percentage of total ESTs (*X*-axis). Abundantly expressed genes in each organism were identified, and the 25 most abundant assemblies were subjected to further functional classification by comparison of the top hit to the Conserved Orthologous Gene list.

Stage-Specific Expression

ESTs from the following nonsubtracted libraries representing different life cycle stages were used to identify stage-specific genes. Tachyzoite/merozoite: TgME49 tachyzoite cDNA, TgRH*-tachyzoite cDNA, TgVEG118 tachyzoite cDNA, TgVEG-tachyzoite cDNA, TgRH tachyzoite cDNA, *P. falciparum* falciparum 3D7 asexual cDNA, *E. tenella* M5–6 merozoite stage. Oocyst/sporozoite: TgVEG partially sporulated oocyst cDNA, TgVEG fully sporulated oocyst cDNA, *E. tenella* S5–2 sporozoite stage. Gametocyte: *P. falciparum* falciparum 3D7 gametocyte cDNA. Bradyzoite: TgME49 invivo bradyzoite cDNA size-selected. For each organism, assemblies with $N \geq 5$ (N = total number of contained ESTs from libraries listed above) were analyzed. Two different strategies were used to identify assemblies that are specifically expressed at a certain stage. One strategy required an assembly to have fivefold more ESTs from one stage than the number present in any other stage (after normalizing its EST content by the ratio of the total sample sizes from different stages). The other strategy required an assembly to have a statistically significantly larger number of ESTs from one stage compared with the other stages. A binomial distribution was assumed for the number of ESTs from a given stage for an assembly. The null hypothesis is that if there is no difference in abundance between stages, the frequency of ESTs in each life cycle stage for an assembly will equal the percentage of total ESTs represented by this stage. To account for multiple testing, a Bonferroni correction was made for p values calculated for each assembly, which was done by multiplying them by the total number of assemblies included for this organism. The cutoff of corrected $p \leq 0.5$ was used for reporting.

Identification of SNPs

Putative SNPs are highlighted in the displays of assembly alignments by calculating the fraction of the dominant nucleotide in columns containing more than 5 characters. If the fraction of the dominant nucleotide falls below 0.8 and both the dominant and second most abundant nucleotides are not a gap or *N* characters, then this column is considered a putative SNP and is highlighted in red to aid identification by users.

Phylogenetic Analyses

Alignments of small subunit rRNA sequences were generated by CLUSTALW (Higgins et al. 1996; http://www.ebi.ac.uk/clustalw/) using default settings and adjustments to correspond to a structural alignment. Structural information was obtained from the comparative RNA Web site (Cannone et al. 2002; http://www.biomedcentral.com/1471-2105/3/2). The resulting alignment had a length of 2427 positions; however, 792 positions were excluded from subsequent analyses because they could not be confidently aligned. Two types of phylogenetic analyses, including 1000 bootstrap replicates each, were performed: parsimony and distance. Parsimony analysis was performed using a heuristic search with 10,000 random stepwise additions. One island of two trees was found. Distances were calculated using the HKY85 model and an among-site rate variation with a γ distribution shape of 0.5. The tree was constructed using neighbor-joining. The only differences between the trees generated by these two methods were the relationships among the *Plasmodium* species. Otherwise, the trees were identical. Only the neighbor-joining distance tree is shown. All phylogenetic analyses were performed with PAUP*4.0 (Swofford 1998). GenBank accession numbers for the sequences are as follows: L19077, *Babesia bovis*; ABO32434, *Babesia microti*; M64243, *Theileria annulata*; L02366, *Theileria parva*; AF112569, *Cryptosporidium parvum*; U07812, *S. neurona*; AF026388, *E. tenella*; U17346, *N. caninum*; X75453, *T. gondii*; M19712, *Plasmodium berghei*; AF180727, *Plasmodium yoelii*; M19172, *Plasmodium falciparum*; Z25819, *Plasmodium reichenowi*; M61723, *Plasmodium gallinaceum*; L07560, *Plasmodium knowlesi*; U03079, *Plasmodium vivax*; M14649, *Procentrum micans*; M88521, *Symbiodinium microadriaticum*; X03772, *Parameceum tetraurelia*; M14601, *Oxytricha nova*. Within the plasmodia, only type A ribosomal genes were included.

ACKNOWLEDGMENTS

Support for the sequencing was provided by the USDA (NRI 99-35204-8615 to L.D.S.), the NIH (AI45806, AI45806A1, AI48121 to L.D.S.), the Amerman Family Foundation (to D.K.H. at the University of Kentucky), Merck Research Laboratories, and the Burroughs Wellcome Fund (to C.S. and D.S.R.). We thank Jonathan Crabtree for providing tools for Web interface development and Gregory Grant for advice on statistical analyses. We are grateful to members of the Washington University Genome Sequencing Center EST, Processing and Annotation Teams, members of Penn Center for Bioinformatics, and members of the Sibley and Roos laboratories who provided advice and assistance. We also thank Jim Ajioka and Depobam Chakrabarti for providing cDNA libraries, Jane Weismann for expert assistance with data submissions at NCBI, and Michael Gottlieb for his untiring support of parasite genomics.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Ajioka, J.A., Boothroyd, J.C., Brunk, B.P., Hehl, A., Hillier, L., Manger, I.D., Overton, G.C., Marra, M., Roos, D., Wan, K.L., et al. 1998. Gene discovery by EST sequencing in *Toxoplasma gondii* reveals sequences restricted to the Apicomplexa. *Genome Res.* **8:** 18–28.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Asai, T., Miura, S., Sibley, L.D., Okabayashi, H., and Takeuchi, T. 1995. Biochemical and molecular characterization of nucleoside triphosphate hydrolase isozymes from the parasitic protozoan *Toxoplasma gondii*. *J. Biol. Chem.* **270:** 11391–11397.

Asai, T., Howe, D.K., Nakajima, K., Nozaki, T., Takeuchi, T., and Sibley, L.D. 1998. *Neospora caninum*: Tachyzoites express a potent Type-I nucleoside triphosphate hydrolase, but lack nucleoside diphosphate hydrolase activity. *Exp. Parasitol.* **90:** 277–285.

Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., and Doolittle, W.F. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290:** 972–977.

Boothroyd, J.C., Hehl, A., Knoll, L.J., and Manger, I.D. 1997. The surface of *Toxoplasma*: More and less. *Intl. J. Parasitol.* **28:** 3–9.

Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D’Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Muller, K.M., et al. 2002. The comparative RNA Web (CRW) site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BioMed Central Bioinformatics* **3:** 2.

Carlton, J.M., Angiuoli, S.V., Suh, B.B., Kooij, T.W., Pertea, M., Silva, J.C., Ermolaeva, M.D., Allen, J.E., Selengut, J.D., Koo, H.L., et al. 2002. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelli yoelli*. *Nat. (Lond.)*

419: 512–519.

Carruthers, V.B. and Sibley, L.D. 1997. Sequential protein secretion from three distinct organelles of *Toxoplasma gondii* accompanies invasion of human fibroblasts. *Eur. J. Cell Biol.* **73:** 114–123.

Chakrabarti, D., Reddy, G.R., Dame, J.B., Almira, E.C., Lapis, P.J., Ferl, R.J., Yang, T.P., Rowe, T.C., and Schuster, S.M. 1994. Analysis of expressed sequence tags from *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **66:** 97–104.

Clark, M.D., Hennig, S., Herwig, R., Clifton, S.W., Marra, M.A., Lehrach, H., and Johnson, S.L. 2001. An oligonucleotide fingerprint normalized and expressed sequence tag characterized zebrafish cDNA library. *Genome Res.* **11:** 1594–1602.

Cummings, C.A. and Relman, D.A. 2000. Using DNA microarrays to study host–microbe interactions. *Emerg. Infect. Dis.* **6:** 513–525.

Davidson, S.B., Crabtree, J., Brunk, B.P., Schug, J., Tannen, V., Overton, G.C., and Stoekert, J. 2001. K2/Kleisi and GUS: Experiments in integrated access to genomic data sources. *IBM Systems J.* **40:** 512–531.

Donahue, C.G., Carruthers, V.B., Gilk, S.D., and Ward, G.E. 2000. The *Toxoplasma* homolog of *Plasmodium* apical membrane antigen-1 (AMA-1) is a microneme protein secreted in response to elevated calcium levels. *Mol. Biochem. Parasitol.* **111:** 15–30.

Dubey, J.P. 1977. *Toxoplasma*, *Hammondia*, *Besniotia*, *Sarcocystis,* and other tissue cyst-forming coccidia of man and animals. In *Parasitic protozoa* (ed. J.P. Kreier), pp. 101–237. Academic Press, New York, NY.

Dubey, J.P. and Lindsay, D.S. 1996. A review of *Neospora caninum* and neosporosis. *Vet. Parasitol.* **67:** 1–59.

Dubremetz, J.F., Achbarou, A., Bermudes, D., and Joiner, K.A. 1993. Kinetics and pattern of organelle exocytosis during *Toxoplasma gondii* host–cell interaction. *Parasitol. Res.* **79:** 402–408.

Dzierszinski, F., Mortuaire, M., Dendouga, N., Popescu, O., and Tomavo, S. 2001. Differential expression of two plant-like enolases with distinct enzymatic and antigenic properties during stage conversion of the protozoan parasite *Toxoplasma gondii*. *J. Mol. Biol.* **309:** 1017–1027.

Eisen, M.B. and Brown, P.O. 1999. DNA arrays for analysis of gene expression. *Methods Enzymol.* **303:** 179–205.

Escalante, A.A. and Ayala, F.J. 1994. Phylogeny of the malarial genus *Plasmodium*, derived from rRNA gene sequences. *Proc. Natl. Acad. Sci.* **91:** 11373–11377.

Ewing, R.M., Kahla, A.B., Poirot, O., Lopez, F., Audic, S., and Claverie, J.M. 1999. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.* **9:** 950–959.

Gajadhar, A.A., Marquardt, W.C., Hall, R., Gunderson, J., Ariztia-Carmona, E.V., and Sogin, M.L. 1991. Ribosomal RNA sequences of *Sarcocystis muris*, *Theileria annulata* and *Crypthecodinium cohnii* reveal evolutionary relationships among Apicomplexans, dinoflagelates, and ciliates. *Mol. Biochem. Parasitol.* **45:** 147–154.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., et al. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature (Lond.)* **419:** 498–511.

Hehl, A.B., Lekutis, C., Grigg, M.E., Bradley, P.J., Dubremetz, J.F., Ortega-Barria, E., and Boothroyd, J.C. 2000. *Toxoplasma gondii* homologue of *Plasmodium* apical membrane antigen 1 is involved in invasion of host cells. *Infect. Immun.* **68:** 7078–7086.

Hettman, C., Herm, A., Geiter, A., Frank, B., Schwarz, E., Soldati, T., and Soldati, D. 2000. A dibasic motif in the tail of a class XIV Apicomplexan myosin is an essential determinant of plasma membrane localization. *Mol. Biol. Cell* **11:** 1385–1400.

Higgins, D.G., Thompson, J.D., and Gibson, T.J. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266:** 382–402.

Hillier, L., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Deitrich, N., Dubuque, T., Favello, A., Gish, W., et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6:** 807–828.

Howe, D.K. 2001. Initiation of a *Sarcocystis neurona* expressed sequence tag (EST) sequencing project: A preliminary report. *Vet. Parasitol.* **95:** 233–239.

Howe, D.K. and Sibley, L.D. 1995. *Toxoplasma gondii* comprises three clonal lineages: Correlation of parasite genotype with human disease. *J. Infect. Dis.* **172:** 1561–1566.

Kissinger, J.A., Brunk, B.P., Crabtree, J., Fraunholz, M.J., Gajria, B., Milgram, A.J., Pearson, D.S., Schug, J., Bahl, A., Diskin, S.J., et al. 2002. The *Plasmodium* genome database. *Nat. (Lond.)* **419:** 490–492.

Kohler, S., Delwiche, C.F., Denny, P.W., Tilney, L.G., Webster, P., Wilson, R.J.M., Palmer, J.D., and Roos, D.S. 1997. A plastid of probable green algal origin in Apicomplexan parasites. *Science* **275:** 1485–1489.

Lashkari, D.A., Derisi, J.L., McCusker, J.H., Namath, A.F., Gentile, C., Hwang, S.Y., Brown, P.O., and Davis, R.W. 1997. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci.* **94:** 13057–13062.

Lekutis, C., Ferguson, D.J., and Boothroyd, J.C. 2000. *Toxoplasma gondii*: Identification of a developmentally regulated family of genes related to SAG2. *Exp. Parasitol.* **96:** 89–96.

Levine, N.D. 1970. Taxonomy of the sporozoa. *J. Protozool.* **56:** 208–209.

Lindsay, D.S. and Dubey, J.P. 1989. In vitro development of *Neospora caninum* (Protozoa: Apicomplexa) from dogs. *J. Parasitol.* **75:** 163–165.

Long, P. 1993. Avian coccidiosis. In *Parasitic protozoa* (ed. J.P. Krier), pp. 1–75. Academic Press, New York, NY.

Manger, I., Hehl, A.B., and Boothroyd, J.C. 1998a. The surface of *Toxoplasma* tachyzoites is dominated by a family of glycosylphosphatidylinositol-anchored antigens related to SAG1. *Infect. Immun.* **66:** 2237–2244.

Manger, I.D., Adrian, H., Parmley, S., Sibley, L.D., Marra, M., Hillier, L., Waterston, R., and Boothroyd, J.C. 1998b. Expressed sequence tag analysis of the bradyzoite stage of *Toxoplasma gondii*: Identification of developmentally regulated genes. *Infect. Immun.* **66:** 1632–1637.

Marra, M.A., Kucaba, T.A., Hillier, L.W., and Waterston, R.H. 1999a. High-throughput plasmid DNA purification for 3 cents per sample. *Nucleic Acids Res.* **27:** e37.

Marra, M., Hillier, L., Kucaba, T., Allen, M., Barstead, R., Beck, C., Blistain, A., Bonaldo, M., Bowers, Y., Bowles, L., et al. 1999b. An encyclopedia of mouse genes. *Nat. Genet.* **21:** 191–194.

Peterson, M.G., Marshall, V.M., Smythe, J.A., Crewther, P.E., Lew, A., Silva, A., Anders, R.F., and Kemp, D.J. 1989. Integral membrane protein located in the apical complex of *Plasmodium falciparum*. *Mol. Cell. Biol.* **9:** 3151–3154.

Pinder, J.C., Fowler, R.E., Dluzewski, A.R., Bannister, L.H., Lavin, F.M., Mitchell, G.H., Wilson, R.J.M., and Gratzer, W.B. 1998. Actomyosin motor in the merozoite of the malaria parasite, *Plasmodium falciparum*, implications for red cell invasion. *J. Cell Sci.* **111:** 1831–1839.

Reddy, G.R., Chakrabarti, D., Schuster, S.M., Ferl, R.J., Almira, E.C., and Dame, J.B. 1993. Gene sequence tags from *Plasmodium falciparum* genomic fragments prepared by the "genase" activity of mung bean nuclease. *Proc. Natl. Acad. Sci.* **90:** 9867–9871.

Roos, D.S., Donald, R.G.K., Morrissette, N.S., and Moulton, A.L. 1994. Molecular tools for genetic dissection of the protozoan parasite *Toxoplasma gondii*. *Methods Cell Biol.* **45:** 28–61.

Roos, D.S., Crawford, M.J., Donald, R.G.K., Kissinger, J.C., Klimczak, L.J., and Striepen, B. 1999. Origin, targeting, and function of the Apicomplexan plastid. *Curr. Opin. Microbiol.* **2:** 426–432.

Scheetz, T.E., Raymond, M.R., Nishimura, D.Y., McClain, A., Roberts, C.W., Birkett, C., Gardiner, J., Zhang, J., Butters, N., Sun, C., et al. 2001. Generation of a high-density rat EST map. *Genome Res.* **11:** 497–502.

Schug, J., Diskin, S., Mazzarelli, J., Brunk, B.P., and Stoeckert Jr., C.J. 2002. Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Res.* **12:** 648–655.

Sibley, L.D., Håkansson, S., and Carruthers, V.B. 1998. Gliding motility: An efficient mechanism for cell penetration. *Curr. Biol.* **8:** R12–R14.

Soldati, D., Dubremetz, J.F., and Lebrun, M. 2001. Microneme proteins: Structural and functional requirements to promote adhesion and invasion by the Apicomplexan parasite *Toxoplasma gondii*. *Intl. J. Parasitol.* **31:** 1293–1302.

Swofford, D.L. 1998. *PAUP phylogenetic analysis using parsimony and other methods. Version 4.* Sinauer Associates, Sunderland, MA.

Tilley, M., Fishera, M., Jerome, M.E., Roos, D.S., and White, M.W. 1997. *Toxoplasma gondii* sporozoites form a transient vacuole that is impermeable and contains only a subset of dense granule proteins. *Infect. Immun.* **65:** 4598–4605.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. *Science* **270:** 484–487.

Wan, K.L., Blackwell, J.M., and Ajioka, J.W. 1995. *Toxoplasma gondii* expressed sequence tags: Insight into tachyzoite gene expression. *Mol. Biochem. Parasitol.* **75:** 179–186.

Waters, A.P., Thomas, A.W., Deans, J.A., Mitchell, G.H., Hudson, D.E., Miller, L.H., McCutchan, T.F., and Cohen, S. 1990. A merozoite receptor protein from *Plasmodium knowlesi* is highly conserved and distributed throughout *Plasmodium. J. Biol. Chem.* **265:** 17974–17979.

WEB SITE REFERENCES

http://blast.wustl.edu; BLAST, Washington University.
http://genome.wustl.edu/EST/; EST databases at Washington University.
http://PlasmoDB.org/; *Plasmodium* genome database, University of Pennsylvania.
http://prodes.toulouse.inra.fr/prodom/doc/prodom.html; Protein Domain ProDom database.
http://www.allgenes.org; All Genes database, Jonathan Crabtree.
http://www.biomedcentral.com/1471-2105/3/2; RNA structure database.
http://www.cbil.upenn.edu/downloads/DoTS/AllGenesRelease4.0/; AllGenes database, The Computational Biology and Informatics Laboratory, University of Pennsylvania.
http://www.cbil.upenn.edu/paradbs-servlet/; ApiESTDB database, University of Pennsylvania.
http://www.cbil.upenn.edu/paradbs-servlet/pub/SupTable1Top25.htm; Supplemental data Table 1, University of Pennsylvania.
http://www.cbil.upenn.edu/paradbs-servlet/pub/SupTable2Stage.htm; Supplemental data Table 2, University of Pennsylvania.
http://www.cbil.upenn.edu/paradbs-servlet/pub/SupTable3Api.htm; Supplemental data Table 3, University of Pennsylvania.
http://www.cbil.upenn.edu/paradbs-servlet/pub/SupTable4Api/; Supplemental data Table 4, University of Pennsylvania.
http://www.ebi.ac.uk/clustalw/; ClustalW alignment, European Biotechnology Institute.
http://www.nature.com/nature/malaria/index.html; *Nature* publishing Web site for the recent publication of malarial genomes.
http://www.ncbi.nlm.nih.gov/cgi-bin/COG/; Conserved orthologous genes, NCBI.
http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml; Conserved domain database, NCBI.
http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/; Taxonomic classifications, NCBI.
http://www.niaid.nih.gov/dmid/malaria/malrep/; Malaria reagent repository MR4, NIH.
http://www.paracel.com/publications/cap4_092200.pdf; CAP4 algorithm, Paracel.

Received August 5, 2002; accepted in revised form December 6, 2002.