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

Follow this and additional works at: [https://digitalcommons.wustl.edu/open_access_pubs](https://digitalcommons.wustl.edu/open_access_pubs)  
Please let us know how this document benefits you.

**Recommended Citation**  
Pape, Deana; Martin, John; Wylie, Todd; Dante, Mike; Waterston, Robert H.; Clifton, Sandra W.; and et al,  
[https://digitalcommons.wustl.edu/open_access_pubs/2111](https://digitalcommons.wustl.edu/open_access_pubs/2111)

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
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

References

This article cites 47 articles, 23 of which can be accessed free at: http://genome.cshlp.org/content/13/3/443.full.html#ref-list-1

Creative Commons License

This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.
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

1Department of Biology, 2Center for Bioinformatics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; 3Department of Genetics, University of Georgia, Athens, Georgia 30602, USA; 4Genome Sequencing Center, Department of Genetics, 5Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, Missouri 63108, USA; 6Department of Veterinary Sciences, University of Kentucky, Lexington, Kentucky 40546, USA; 7Human and Animal Infectious Diseases, Merck Research Laboratories, Rahway, New Jersey 07065, USA; 8Veterinary 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 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
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 ancestor 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 under way. 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’ ends 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](https://www.genome.org)
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 < 10⁻⁰⁵. 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 (<30%) of ESTs actually extend into the 5'-UTR region in *T. gondii* genes (Ajioka et al. 1998; Manger et al. 1999b). 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
were quite selective, and it is likely that more sensitive searches would identify additional similarities. ApisESTDB 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. S), 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. S) 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.
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 \textit{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 \textit{T. gondii} and \textit{N. caninum} were quite similar and were dominated by secretory proteins including those found in rhoptries, dense granules, and micronemes. 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 \textit{E. tenella} and \textit{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 \textit{T. gondii} (Manger et al. 1998a; Boothroyd et al. 1997; Lekutis et al. 2000). Members of this family were identified in \textit{N. caninum} as well as in \textit{S. neurona}, and a related, but distinct family of cysteine-rich surface antigens was abundant in \textit{E. tenella}. Finally, the pattern of highly expressed ESTs in \textit{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 \textit{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-
The availability of an EST gene database will facilitate such studies 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 Table 1 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 number 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 2. Stage-Specific Expression of Gene Families in Apicomplexans**

<table>
<thead>
<tr>
<th>Organism</th>
<th>Tachyzoite/merozoite</th>
<th>Oocyst/sporozoite</th>
<th>Gametocyte</th>
<th>Bradyzoite</th>
<th>Constitutive</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Toxoplasma gondii</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fivefold</td>
<td>234 (41.2)*</td>
<td>66 (11.6)</td>
<td>ND</td>
<td>69 (12.1)</td>
<td>199 (35.0)</td>
</tr>
<tr>
<td><em>P</em> ≤ 0.05</td>
<td>13 (2.3)</td>
<td>48 (8.5)</td>
<td>32 (5.6)</td>
<td>475 (83.6)</td>
<td></td>
</tr>
<tr>
<td><em>Plasmodium falciparum</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fivefold</td>
<td>60 (29.6)</td>
<td>ND</td>
<td>108 (53.2)</td>
<td>NA</td>
<td>35 (17.2)</td>
</tr>
<tr>
<td><em>P</em> ≤ 0.05</td>
<td>14 (6.9)</td>
<td>10 (4.9)</td>
<td>179 (88.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Eimeria tenella</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fivefold</td>
<td>65 (46.4)</td>
<td>51 (36.4)</td>
<td>ND</td>
<td>NA</td>
<td>24 (17.1)</td>
</tr>
<tr>
<td><em>P</em> ≤ 0.05</td>
<td>15 (10.7)</td>
<td>16 (11.4)</td>
<td></td>
<td></td>
<td>109 (77.9)</td>
</tr>
</tbody>
</table>

Based on assemblies where *N* ≥ 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* ≤ 0.05. NA, not applicable; ND, not done.

*Values expressed as total assemblies (% of total).*

---

**Table 3. Closest Phylogenetic Ortholog for Apicomplexan Gene Assemblies**

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

*Total number of assemblies including mRNA singletons.

*Number of assemblies in category (%).*
lications 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^{-5} 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 than to those of their mammalian counterparts; (3) those that are specifically restricted to the Apicomplexa; and (4) those that are more highly conserved within the Apicomplexa, indicating that they fulfill specific roles unique to the members of this phylum.

Several conserved enzymes and signaling molecules common to many taxa are also found on this list, such as the apicomplexan assembly (p-value < 10^{-5}), but were similar to at least another apicomplexan-specific 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 T. 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
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 (Ajoka 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 fibroblasts (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.). cDNAs were synthesized by oligo(dT) priming of poly(A)-tract mRNA, size-selected, and directionally cloned 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 (Stratagene). This library has been deposited with the MR4 collection as entry MRA-101 (at http://www.mr4.org). 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 XLI-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 electroperoration 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 (Robbins Scientific) robot to each well of a 96-well Beckman block, containing 1 mL of Terrific Broth (DiCor) 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 electropherograms (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, S2
(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 andmask (G. Miklén, unpubl.). (G which uses BLASTX (S = 50) to compare with a database of T. gondii repeat elements (GenBank accession nos. MS7916, MS7917, MS7918, MS7919, 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 these 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 E2 = 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), human mito mitochondrial (BLASTN2 S = 170, gap S2 = 150, M = 5, N = -11, Q = 11, R = 11, W = 10, E2 = 0.5 against GenBank hummc and musmc) 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% identity 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 (0.7%). 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 cell 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 mouse gene sequences called WORM (E. Sonnhammer, unpubl.). These identifications 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 to 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.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://prodom.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/scf.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^-3 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.ncbi.nlm.nih.gov/Genomes/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 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 to 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.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.
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, TgVEG-tachyzoite cDNA, P. falciparum falciparum 3D7 asexual cDNA, E. tenella MS-6 merozoite stage. Oocyst/sporozoite: TgME49 partially sporulated oocyst cDNA, TgVEG fully sporulated oocyst cDNA, E. tenella SS-2 sporozoite stage. Gametocyte: P. falciparum falciparum 3D7 gametocyte cDNA. Bradyzoite: TgME49 invivo bradyzoite cDNA size-selected. For each organism, assemblies with N ≥ 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 there is no difference in abundance between stages, the frequency of ESTs in each life cycle stage for an assembly will equal the frequency of the top hit for the top hit to the Conserved Orthologous Gene list.

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/1471-2105/3/12). 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; AFO26388, E. tenella; U17346, N. caninum; X75453, T. gondii; M19712, Plasmodium berghei; AF180727, Plasmodium yoelii; M19172, Plasmodium falciparum; Z22819, Plasmodium reichenowi; M61723, Plasmodium gallinaceum; L07560, Plasmodium knowlesi; U03079, Plasmodium vivax; M14649, Procentrum micans; M88521, Symbo- odinium microadriaticum; X035772, Paramecium 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, AI45806A, AI48121 to L.D.S.), the American Family Foundation (to D.K.H. at the University of Kentucky), Merck Research Laboratories, and the Burroughs Wellcome Fund (to C.S. and O.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 Ajikoka 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


**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://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/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.niaid.nih.gov/dmid/malaria/malrep/; Malaria reagent repository MR4, NIH.


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