Database Tool

The *Eimeria* Transcript DB: an integrated resource for annotated transcripts of protozoan parasites of the genus *Eimeria*

Luiz Thibério Rangel[1], Jeniffer Novaes[1], Alan M. Durham[2], Alda Maria B. N. Madeira[1] and Arthur Gruber[1],*

[1]Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, Avenida Professor Lineu Prestes 1374, São Paulo SP 05508-000, Brazil and [2]Department of Computer Sciences, Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão 1010, Bloco C, São Paulo SP 05508-000, Brazil

*Corresponding author: Tel: +55 11 3091 7274; Fax: +55 11 3091 7417; Email: argruber@usp.br

Parasites of the genus *Eimeria* infect a wide range of vertebrate hosts, including chickens. We have recently reported a comparative analysis of the transcriptomes of *Eimeria acervulina*, *Eimeria maxima* and *Eimeria tenella*, integrating ORESTES data produced by our group and publicly available Expressed Sequence Tags (ESTs). All cDNA reads have been assembled, and the reconstructed transcripts have been submitted to a comprehensive functional annotation pipeline. Additional studies included orthology assignment across apicomplexan parasites and clustering analyses of gene expression profiles among different developmental stages of the parasites. To make all this body of information publicly available, we constructed the *Eimeria* Transcript Database (*Eimeria*TDB), a web repository that provides access to sequence data, annotation and comparative analyses. Here, we describe the web interface, available sequence data sets and query tools implemented on the site. The main goal of this work is to offer a public repository of sequence and functional annotation data of reconstructed transcripts of parasites of the genus *Eimeria*. We believe that *Eimeria*TDB will represent a valuable and complementary resource for the *Eimeria* scientific community and for those researchers interested in comparative genomics of apicomplexan parasites.

Database URL: http://www.coccidia.icb.usp.br/eimeriatdb/

Introduction

Coccidian parasites infect a wide range of vertebrate hosts and cause many diseases of veterinary and human importance. Within this group, parasites of the genus *Eimeria* infect many species of wild and domestic hosts, including poultry. Seven distinct *Eimeria* species may infect chickens, causing enteric diseases that lead to diarrhoea, malabsorption, impaired weight gain and higher susceptibility to opportunistic diseases. The economic impact of such diseases is reflected by direct costs associated with the reduced productivity in affected flocks and by indirect costs related to the preventive use of anti-coccidial drugs and/or vaccines (1). The production losses because of coccidiosis have been estimated at US$ 2400 million per annum worldwide (2). *Eimeria* parasites are easily propagated through oral infections of experimental animals under controlled conditions, thus permitting to perform molecular studies. The genome size of *Eimeria tenella*, the model species, comprises ∼55–60 Mb distributed in 14 chromosomes (3). The complete sequence of *E. tenella* chromosome 1 (4) and a whole-genome sequence are publicly available on GeneDB (5) and EuPathDB (6) databases. In addition, a draft sequence of the *Eimeria maxima* genome has been recently reported (7). The transcriptome of *Eimeria*

parasites has been assessed using conventional Expressed Sequence Tag (EST) (8–13) and ORESTES (open reading frame ESTs) reads (14). In addition, Amiruddin *et al.* (15) obtained full-length cDNA sequences of 443 *E. tenella* genes. The cDNA libraries used in all these works have been derived mainly from the most accessible developmental stages, including oocysts in different phases of sporulation, sporozoites and first- and second-generation merozoites. Based on genomic gene prediction and transcriptome assembly data, the transcriptome complexity of *E. tenella* has been estimated as circa 8700 genes (14). We have recently reported an integrated and comparative analysis of the transcriptome of *Eimeria acervulina*, *E. maxima* and *E. tenella*, including ORESTES data produced by our group and publicly available ESTs (14). All cDNA reads were assembled and the reconstructed transcripts submitted to a comprehensive functional annotation pipeline. Comparative studies included orthology assignment across apicomplexan parasites and clustering analyses of gene expression profiles among different developmental stages of the parasites. To make all this body of information freely available to the scientific community, we constructed the *Eimeria* Transcript Database (*Eimeria*TDB), a web repository that provides access to sequence data, annotation and comparative analyses. Here, we describe the web interface, available sequence data sets and query tools implemented on the site. The main goal of this work is to offer a public repository of sequence and functional annotation data of reconstructed transcripts of parasites of the genus *Eimeria*.

Data content of current release

*Eimeria*TDB v. 1.1 contains transcript sequences of *E. acervulina*, *E. maxima* and *E. tenella*, reconstructed from EST and ORESTES data, as previously described (14). In total, the current version comprises data sets of 3413, 3426 and 8700 assembled transcripts, respectively. The cDNA reads are derived from several developmental stages of the parasites, including unsporulated oocysts, sporoblast-phase oocysts, sporulated oocysts, sporozoites and first- and second-generation merozoites. *Eimeria*TDB comprises assembled and unassembled data, annotation of individual assembled sequences and global analysis of each transcriptome data set. Digital expression data, based on the frequency of reads belonging to each assembled transcript, are also available.

Database organization and implementation

The assembled transcripts of the three *Eimeria* species were submitted to an annotation pipeline constructed with EGene2, a new version of the platform (16) that includes annotation components (available on request). Briefly, the pipeline consisted in finding all potential ORFs and translating into the corresponding products. We used an arbitrarily chosen ORF length of at least 50 amino acids. All protein products were inspected for sequence similarity using BLASTp (17) against the NCBI non-redundant protein database, protein domains using RPS-BLAST against Conserved Domains Database (CDD) (18), protein motifs with InterProScan (19), signal peptide and transmembrane domain prediction using Phobius (20) and Glycosylphosphatidyl inositol (GPI) anchoring cleavage site prediction using DGPI (Kronegg and Buloz, unpublished results, downloaded from http://129.194.185.165/dgpi/ on March 2008). Finally, using InterPro IDs, we mapped and quantified Gene Ontology (GO) terms (21) using a GO slim file, a subset of GO terms. Also, all proteins were functionally classified using KOG (22) and eggNOG (23) databases of orthology and mapped onto the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway (24) database. We also performed an integrated orthology analysis of the translated products of the three *Eimeria* species with data sets of proteins predicted from genomes of six apicomplexan parasites: *Toxoplasma gondii*, *Plasmodium falciparum*, *Neospora caninum*, *Babesia bovis*, *Theileria annulata* and *Cryptosporidium parvum*. For this task, we used the programs InParanoid (25) and MultiParanoid (26), as previously described (14). The extensible mark-up language annotation file generated by EGene was used to automatically populate a MySQL database using an in-house script. The web interface was developed using PHP, HTML and JavaScript languages, and it was integrated with the database through a set of in-house Perl scripts. All annotation data were integrated with the Generic Genome Browser (GBrowse), a genome viewer that is widely used for visualization of sequence annotation (27). *Eimeria*TDB is linked to the NCBI BioProject Database under the accession codes PRJNA81161, PRJNA81163 and PRJNA81165. The repository is publicly available at http://www.coccidia.icb.usp.br/eimeriatdb/. Publications that use this database should cite the aforementioned URL and this publication.

Data analysis tools

*Eimeria*TDB offers a variety of services, including a local BLAST engine, a database-querying page, annotation pages of individual transcripts, global analyses of whole data sets and a data download page. Table 1 depicts all resources provided in the website and the respective descriptions. The interface presents a set of tabs at the main page, each one redirecting to a specific service page (Figure 1A).

BLAST

A local BLAST service is available, and searches can be performed against many *Eimeria* databases, including

Table 1. Resources available at the *Eimeria* Transcript Database (*Eimeria*TDB) web site

Resource	Features
Online analysis tools	Local BLAST engine Similarity searches against different *E. tenella* genome assembly versions, assembled and unassembled EST/ORESTES of *E. acervulina*, *E. maxima* and *E. tenella*
Relational database	Queries using sequence IDs, keywords and evidence results. Access to annotation results of individual sequences
Annotation	
Individual annotation	Evidence Orthology analysis with other apicomplexans, BLAST x nr, RPS-BLAST x CDD, InterproScan searches, SignalP, TMHMM, Phobius, DGPI analyses, GO term mapping, functional classification using KOG and eggNOG and pathway mapping using KEGG. Expression analysis, when available, displayed in a graphic Downloads DNA transcript sequence, DNA ORF sequences and protein sequences Annotation reports All sequences annotated with or without ORF selection. Available formats: Feature Table, Feature Table with Artemis additional feature keys and GFF3. Graphical visualization of annotation data on GBrowse.
Global annotation	GO term mapping Classification and quantification of assembled cDNAs into GO terms Orthology analysis Classification and quantification of assembled cDNAs into functional groups of KOG and eggNOG Pathway mapping Classification and quantification of assembled cDNAs into KEGG's Metabolic Pathway Classes
Downloads	cDNA products (ORFs >50), assembled cDNA sequences, annotation reports in Feature Table, Feature table with Artemis additional feature keys and GFF3 formats.

genomic, cDNA and mitochondrial sequences. Genomic sequences comprise shotgun reads and several assembly versions from the Wellcome Trust Sanger Institute (ftp://ftp.sanger.ac.uk/pub/pathogens/Eimeria/tenella/). Expressed sequences include assembled cDNAs of *E. tenella*, *E. acervulina* and *E. maxima*. In the case of *E. tenella*, the database contains an assembly constructed from a mixture of ORESTES and EST reads (14). For *E. acervulina* and *E. maxima*, the current version of the database contains assemblies obtained with ORESTES reads only. All programs of BLAST package can be used: blastn, blastp, blastx, tblastx and tblastn. Once a given assembled cDNA hit is identified, the user can consult the relational database to inspect the corresponding annotation using the sequence ID.

Querying the database

The Search Database section allows users to perform customized queries to *Eimeria*TDB (Figure 1B). The database integrates data from the three *Eimeria* species and results from all programs used to collect evidence. If the user already knows the sequence ID, for example 'Eten_0011', then the corresponding annotation can be directly retrieved. Searches can also be performed using single or multiple query terms. Query terms include product names

(e.g. hexokinase, serine protease, microneme protein and so forth), descriptions and IDs derived from InterPro, KOG (e.g. KOG1696; 60S ribosomal protein L19), eggNOG (e.g. euNOG10377; transporter protein) and KEGG (e.g. citrate cycle; K01647, citrate synthase; large subunit ribosomal protein L19e and so forth). Queries can be restricted using different sets of radio buttons to a specific *Eimeria* species, or according to different types of evidence. In the latter case, search results can be restricted to only those sequences presenting a given subset of results. For instance, a user can specify 'receptor' as a keyword and restrict the results to the sequences presenting positive results for transmembrane domains and signal peptide. In this case, the sequences retrieved by the search are most probably related to membrane bound proteins, such as G-protein coupled and other receptors. As a result of the query, the user obtains a list of sequences fulfilling the search criteria, with specific links to the respective annotation pages.

Exploring transcript annotation

Annotation is provided in three distinct formats: Feature Table (FT), extended FT and Generic Feature Format (GFF) 3 (Figure 1C). FT is the annotation format and vocabulary
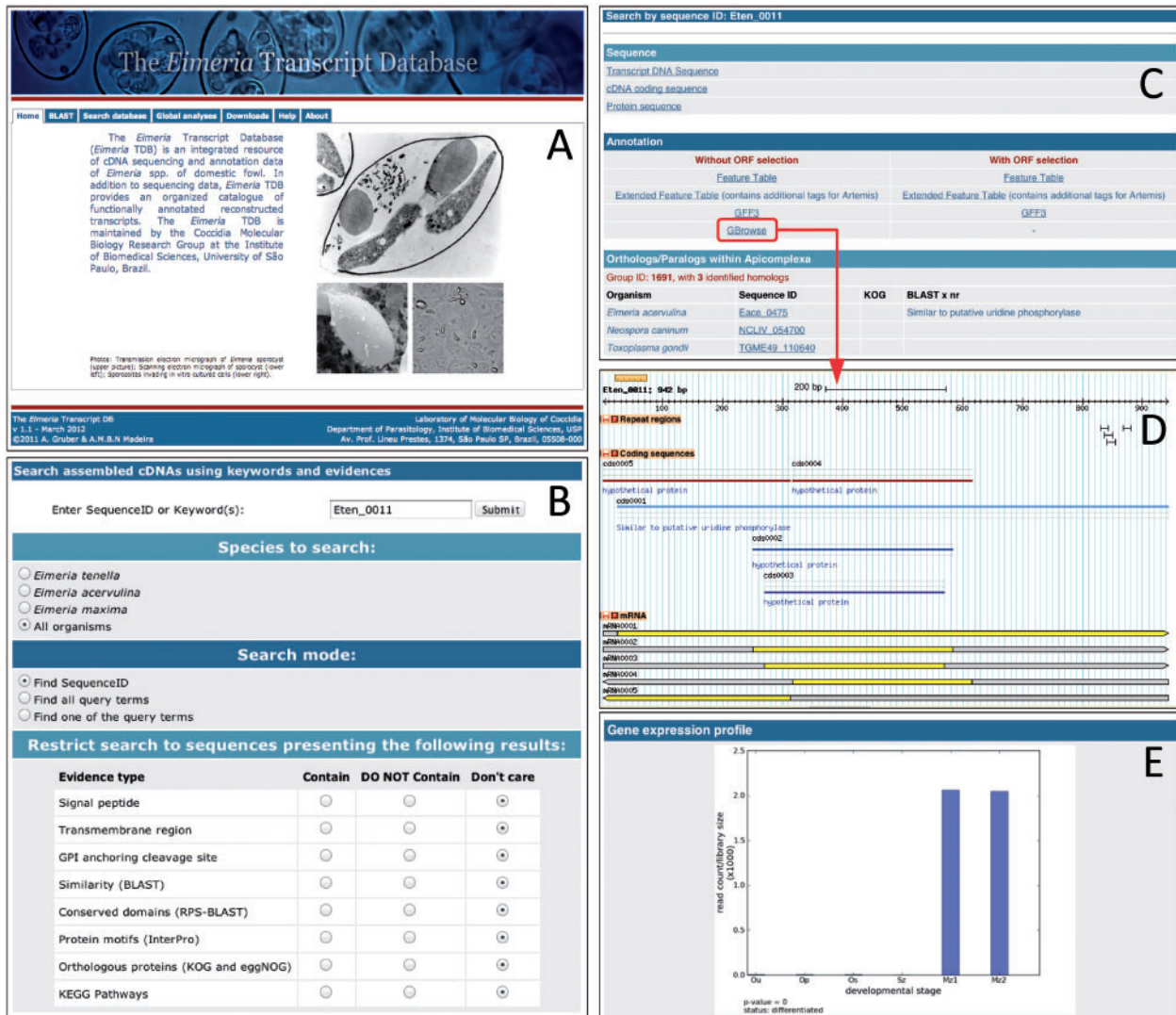
Figure 1. Screenshots of some resources available at *EimeriaTDB*. The home page (**A**) contains tabs redirecting to specific service pages. The search page (**B**) allows querying the database using sequence IDs or keywords. Queries can be restricted to specific *Eimeria* species or according to different types of evidence. The annotation page (**C**) provides access to sequence and annotation data, orthology analysis within apicomplexan organisms and a link to the respective GBrowse (**D**) screen. When available, expression data are displayed in a graphic (**E**).

terms adopted by the main sequence repositories (DDBJ/EMBL/GenBank). A definition of FT is available at the International Nucleotide Sequence Database Collaboration site (http://www.insdc.org/files/feature_table.html). We also provide an extended FT version, which includes some specific tags that are not officially included in the FT specification, but they are compatible with Artemis annotation and editing tool (28). For GFF3, we followed the definition available at the Sequence Ontology Project (http://www.sequenceontology.org/gff3.shtml). The annotation files are available with and without automatic ORF selection (see 'Evidence Annotation' section), and all results (selected and unselected

ORFs) are available for inspection. Also, annotation can be graphically visualized with GBrowse through an available link (Figure 1D).

Orthology analysis across apicomplexan parasites

Orthologues identified in other Apicomplexa organisms are listed in a specific table (Figure 1C), which displays the corresponding sequence IDs, KOG IDs and, when available, BLAST hits. Also, links to the amino acid sequences of the

orthologues are provided. Orthologues of other *Eimeria* species are cross-referenced through links to the respective annotation page at *Eimeria*TDB.

Gene expression profiles

When available, we provide a chart displaying the expression profile of the gene across different *Eimeria* developmental stages (Figure 1E). The expression data of each stage are based on the normalized number of reads comprising each assembled sequence according to their respective source (29). Briefly, we used in-house scripts (available on request) to convert the CAP3 assembly files into spreadsheets that list the number of reads belonging to each assembled sequence with regard to their respective developmental stage, as previously described by Novaes *et al*. (14). The corresponding *P*-value and status of expression (differentiated/non-differentiated) are also displayed.

Evidence-based annotation

The final part of the annotation page provides descriptions and links for all program results that give support to a function for the putative gene (Figure 2A). By clicking on the respective link, the user is redirected to the specific page of each program result, such as sequence alignments and graphical results. In addition, links to mapped GO terms, functional classification using KOG and eggNOG databases and pathway mapping on KEGG are also presented. When available, KEGG results contain links to the corresponding KEGG Orthology (KO) page on KEGG's site and pathway image (Figure 2B). Stored results are also available for the following programs: BLAST (Figure 2C), RPS-BLAST, InterProScan (Figure 2D), SignalP, TMHMM, Phobius and DGPI. Our annotation pipeline has automatically selected the most probable coding ORF, based on weighted criteria on a set of bioinformatics analysis results for each ORF of an assembled transcript. Nevertheless, if the user wants to inspect the results of all ORFs, we provide a link entitled 'evidence for all predicted ORFs' at the bottom of the annotation page.

Global analyses

A specific section of the site provides both qualitative and quantitative analyses for the whole sets of translated products of *E. acervulina*, *E. maxima* and *E. tenella*. Analyses include GO term mapping, orthology functional classification using KOG and eggNOG databases and pathway mapping using KEGG. All annotated proteins are mapped onto GO terms using a GO slim file comprising a subset of GO terms. The results are presented in a composite table comprising the three ontology domains, with the respective GO slim terms and sequence counts. If the user clicks on the GO

term itself, the page is redirected to the AmiGO browser (30), showing the corresponding term description. Also, there are links to all sequences whose products have been mapped to the particular GO term. All translated protein sequences are also mapped onto KEGG Orthology database, and the corresponding pathways are identified. The KEGG Pathway classes are listed on a table with the respective sequence counts (Figure 2E), and distribution is depicted in a pie chart. By clicking on a KEGG Pathway Class link (e.g. metabolism), an expanded list of subclasses is displayed. Each subclass presents the corresponding number of classified sequences and contains a link that opens up a page with the list of proteins (with links to BLAST alignments), Orthology Group (KO number), KO descriptions, E.C. numbers and KEGG pathways. Each pathway provides a link to the corresponding KEGG pathway image, with the respective query protein highlighted in a red-labelled box (Figure 2B). Finally, the transcript products are also mapped onto KOG and eggNOG databases. In both cases, the results are displayed in a table listing the functional categories and the respective number of sequences classified in each one (Figure 2F). By clicking on the one-letter functional class code, a page displaying a list of all proteins classified in this specific category is presented, with links to the corresponding BLAST alignments.

Retrieving data

Each sequence annotation page provides links to the respective transcript DNA sequence and translated product in FASTA format, plus annotation data in Feature Table and GFF3 formats. Also, the Downloads section allows the user to download tarball compressed files that comprise global data sets for each of the three *Eimeria* species, including nucleotide and amino acid sequences and annotation files.

Future directions

By the time our group had described the transcriptome of three *Eimeria* species that infect chickens, Amiruddin *et al*. (15) reported an initiative of full-length transcript sequencing in *E. tenella*, comprising the entire sequence of 443 *E. tenella* transcripts and corresponding to ~5% of the parasite transcriptome. To our knowledge, some other groups are conducting RNAseq studies in different developmental stages of the parasite. We intend to incorporate all publicly available transcriptome sequence data in future releases of the database, thus providing an increasingly higher coverage of *Eimeria* reconstructed transcripts. We currently use a relatively simple database schema for *Eimeria*TDB. However, a newly developed version of the EGene platform will perform automatic
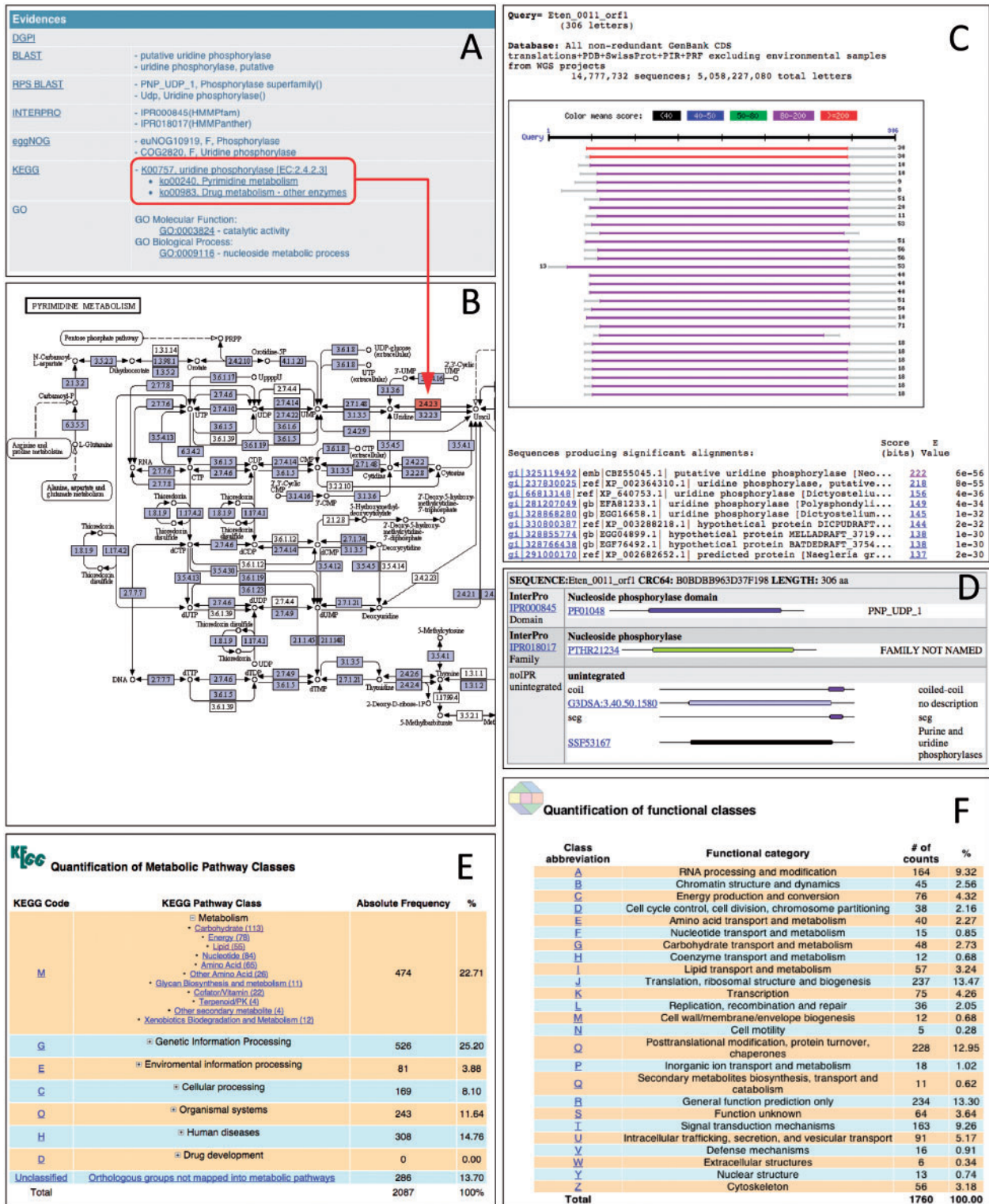
Figure 2. Evidence-based annotation. Each annotation page provides specific links to different bioinformatics analyses (**A**), including sequence mapping onto KEGG pathways (**B**), BLAST similarity (**C**) and InterPro motif searching (**D**). Global analyses include functional classification into KEGG Metabolic Pathway (**E**) and KOG orthology (**F**) classes.

annotation using Chado ([31]), the relational database schema that underlies the Generic Model Organism Database (GMOD) applications. We intend to incorporate this schema and associated annotation in future releases of *Eimeria*TDB.

EuPathDB has recently incorporated genomic and EST data of *E. tenella*, offering a new perspective of comparative analysis across apicomplexan and other protozoan parasites. *Eimeria*TDB, by mainly focusing on transcript reconstruction and annotation, may represent a valuable and complementary resource for the *Eimeria* scientific community, and for those researchers interested in comparative genomics of apicomplexan parasites.

References

1. Williams,R.B. (1999) A compartmentalised model for the estimation of the cost of coccidiosis to the world's chicken production industry. *Int. J. Parasitol.*, **29**, 1209–1229.
2. Shirley,M.W., Smith,A.L. and Tomley,F.M. (2005) The biology of avian *Eimeria* with an emphasis on their control by vaccination. *Adv. Parasitol.*, **60**, 285–330.
3. Shirley,M.W. (2000) The genome of *Eimeria* spp., with special reference to *Eimeria tenella*—a coccidium from the chicken. *Int. J. Parasitol.*, **30**, 485–493.
4. Ling,K.H., Rajandream,M.A., Rivailler,P. *et al*. (2007) Sequencing and analysis of chromosome 1 of *Eimeria tenella* reveals a unique segmental organization. *Genome Res.*, **17**, 311–319.
5. Logan-Klumpler,F.J., De Silva,N., Boehme,U. *et al*. (2012) GeneDB—an annotation database for pathogens. *Nucleic Acids Res.*, **40**, D98–D108.
6. Aurrecoechea,C., Heiges,M., Wang,H. *et al*. (2007) ApiDB: integrated resources for the apicomplexan bioinformatics resource center. *Nucleic Acids Res.*, **35**, D427–D430.
7. Blake,D.P., Alias,H., Billington,K.J. *et al*. (2012) EmaxDB: availability of a first draft genome sequence for the apicomplexan *Eimeria maxima*. *Mol. Biochem. Parasitol.*, **184**, 48–51.
8. Li,L., Brunk,B.P., Kissinger,J.C. *et al*. (2003) Gene discovery in the apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. *Genome Res.*, **13**, 443–454.
9. Schwarz,R.S., Fetterer,R.H., Rosenberg,G.H. *et al*. (2010) Coccidian merozoite transcriptome analysis from *Eimeria maxima* in comparison to *Eimeria tenella* and *Eimeria acervulina*. *J. Parasitol.*, **96**, 49–57.
10. Miska,K.B., Fetterer,R.H. and Barfield,R.C. (2004) Analysis of transcripts expressed by *Eimeria tenella* oocysts using subtractive hybridization methods. *J. Parasitol.*, **90**, 1245–1252.
11. Miska,K.B., Fetterer,R.H. and Rosenberg,G.H. (2008) Analysis of transcripts from intracellular stages of *Eimeria acervulina* using expressed sequence tags. *J. Parasitol.*, **94**, 462–466.
12. Ng,S.T., Sanusi Jangi,M., Shirley,M.W. *et al*. (2002) Comparative EST analyses provide insights into gene expression in two asexual developmental stages of *Eimeria tenella*. *Exp. Parasitol.*, **101**, 168–173.
13. Wan,K.L., Chong,S.P., Ng,S.T. *et al*. (1999) A survey of genes in *Eimeria tenella* merozoites by EST sequencing. *Int. J. Parasitol.*, **29**, 1885–1892.
14. Novaes,J., Rangel,L.T., Ferro,M. *et al*. (2012) A comparative transcriptome analysis reveals expression profiles conserved across three *Eimeria* spp. of domestic fowl and associated with multiple developmental stages. *Int. J. Parasitol.*, **42**, 39–48.
15. Amiruddin,N., Lee,X.W., Blake,D.P. *et al*. (2012) Characterisation of full-length cDNA sequences provides insights into the *Eimeria tenella* transcriptome. *BMC Genomics*, **13**, 21.
16. Durham,A.M., Kashiwabara,A.Y., Matsunaga,F.T. *et al*. (2005) EGene: a configurable pipeline generation system for automated sequence analysis. *Bioinformatics*, **21**, 2812–2813.
17. Altschul,S.F., Madden,T.L., Schaffer,A.A. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
18. Marchler-Bauer,A., Lu,S., Anderson,J.B. *et al*. (2011) CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.*, **39**, D225–D229.
19. Mulder,N. and Apweiler,R. (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol. Biol.*, **396**, 59–70.
20. Kall,L., Krogh,A. and Sonnhammer,E.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
21. Ashburner,M., Ball,C.A., Blake,J.A. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
22. Tatusov,R.L., Fedorova,N.D., Jackson,J.D. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
23. Muller,J., Szklarczyk,D., Julien,P. *et al*. (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.*, **38**, D190–D195.
24. Kanehisa,M., Goto,S., Sato,Y. *et al*. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
25. Ostlund,G., Schmitt,T., Forslund,K. *et al*. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, **38**, D196–D203.

26. Remm,M., Storm,C.E. and Sonnhammer,E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.

27. Stein,L.D., Mungall,C., Shu,S. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–610.

28. Rutherford,K., Parkhill,J., Crook,J. *et al*. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.

29. Audic,S. and Claverie,J.M. (1997) The significance of digital gene expression profiles. *Genome Res.*, **7**, 986–995.

30. Carbon,S., Ireland,A., Mungall,C.J. *et al*. (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.

31. Mungall,C.J., Emmert,D.B. and FlyBase,C. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.