An Orthology-Based Analysis of Pathogenic Protozoa Impacting Global Health: An Improved Comparative Genomics Approach with Prokaryotes and Model Eukaryote Orthologs

Rafael R. C. Cuadrat,^{1*} Sérgio Manuel da Serra Cruz,^{2,3*} Diogo Antônio Tschoeke,^{1*}
Edno Silva,² Frederico Tosta,² Henrique Jucá,¹ Rodrigo Jardim,¹
Maria Luiza M. Campos,⁴ Marta Mattoso,² and Alberto M. R. Dávila¹

Abstract

A key focus in 21st century integrative biology and drug discovery for neglected tropical and other diseases has been the use of BLAST-based computational methods for identification of orthologous groups in pathogenic organisms to discern orthologs, with a view to evaluate similarities and differences among species, and thus allow the transfer of annotation from known/curated proteins to new/non-annotated ones. We used here a profile-based sensitive methodology to identify distant homologs, coupled to the NCBI's COG (Unicellular orthologs) and KOG (Eukaryote orthologs), permitting us to perform comparative genomics analyses on five protozoan genomes. *OrthoSearch* was used in five protozoan proteomes showing that 3901 and 7473 orthologs can be identified by comparison with COG and KOG proteomes, respectively. The core protozoa proteome inferred was 418 Protozoa-COG orthologous groups and 704 Protozoa-KOG orthologous groups: (i) 31.58% (132/418) belongs to the category J (translation, ribosomal structure, and biogenesis), and 9.81% (41/418) to the category O (post-translational modification, protein turnover, chaperones) using COG; (ii) 21.45% (151/704) belongs to the categories J, and 13.92% (98/704) to the O using KOG. The phylogenomic analysis showed four well-supported clades for Eukarya, discriminating Multicellular [(i) human, fly, plant and worm] and Unicellular [(ii) yeast, (iii) fungi, and (iv) protozoa] species. These encouraging results attest to the usefulness of the profile-based methodology for comparative genomics to accelerate semi-automatic re-annotation, especially of the protozoan proteomes. This approach may also lend itself for applications in global health, for example, in the case of novel drug target discovery against pathogenic organisms previously considered difficult to research with traditional drug discovery tools.

Introduction

THE NEGLECTED TROPICAL DISEASES (NTD) affect 1.2 billion people, the so-called “The Bottom Billion,” most of them comprising people who live on less than \$1 per day in the poorest regions of Africa, Asia, Latin America, and the Caribbean (Collier, 2007; Hotez and Brown, 2009). These diseases are classified as a subset of infectious diseases, and the causing pathogens may be viruses, bacteria, protozoa, or helminthes (Feasey et al., 2010). Definitions that classify an NTD may vary; the most comprehensive, defined by Hotez

and Brown (2009), lists 37 NTD, from which six are caused by protozoa.

In recent years, genomes of several parasitic protozoa that cause NTD were sequenced. Various efforts were performed to annotate these genomes, for instance, trypanosomatids as *Trypanosoma cruzi* (El-Sayed et al., 2005a), *Trypanosoma brucei* (Berriman et al., 2005), and *Leishmania major* (Ivens et al., 2005), the so-called Trityp (El-Sayed et al., 2005). Collectively, Trityp cause disease and death of millions of humans and countless infections in other mammals. Thus, the search efforts to develop vaccines and drugs against them are

¹Computational and Systems Biology Laboratory, Computational and Systems Biology Pole, Oswaldo Cruz Institute, Fiocruz, Brazil.

²COPPE/PESC and ⁴PPGI—Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.

³PPGMMC—Federal Rural University of Rio de Janeiro, Seropédica, Brazil.

*These authors contributed equally to this work.

minor when compared to human diseases such as cancer or AIDS (Lindoso and Lindoso, 2009). Two protozoan parasites causing NTD that had their genomes sequenced were *Plasmodium falciparum* and *Entamoeba histolytica*. *P. falciparum* is one of the species of *Plasmodium* that causes malaria in humans. Resistance to current anti-malarial drugs, such as chloroquine, has already been reported (Hastings et al., 2002). *E. histolytica* is the causative agent of human amoebiasis, a cosmopolitan disease whose most common clinical forms are amoebic colitis and amoebic liver abscess (Ximénes et al., 2010).

The Tritryp genome sequencing was completed in 2005 (Berriman et al., 2005; El-Sayed et al., 2005a; Ivens et al., 2005), followed by a comprehensive study comparing the genome architecture of these organisms (El-Sayed et al., 2005b). The study revealed a conserved core proteome corresponding to about 6200 genes in large syntenic polycistronic gene clusters. However, a high percentage (approximately 35%) of those Tritryp genes is non-annotated, which limits the inference of targets for drugs based on the function of these genes (Salavati and Najafabadi, 2010). The same can be said about the genomes of other Protozoa such as *P. falciparum* (~57% annotated as hypothetical proteins) (Brehelin et al., 2010) and *E. histolytica* (~30% annotated as hypothetical proteins) (Lorenzi et al., 2010). In such context, more sensitive or improved methodologies for comparative genomics and consequently re-annotation and/or more curated data from phylogenetically related species might be useful to help with the annotation of these genomes.

Comparative genomics can be used to infer the function of genes and add annotation to new sequences. The basis for this type of analysis is the hypothesis that important biological sequences are conserved between species due to functional constraints. To accomplish biological inferences based on comparative genomics, the first step is the choice of species to be compared. The ideal pairwise comparison is between phylogenetically close organisms (Nobrega and Pennacchio, 2004).

It is desirable to infer biological information and finally obtain knowledge using comparative genomics approaches. While the inference of similarities and differences among genomes/proteomes is the first aim of this kind of approach, improving the annotation of the increasing amount of data originated from the constant influx of new genome sequences may be considered a consequence. It is also a significant goal in creating a comprehensive evolutionary classification of the genes of all sequenced genomes. Such classification is based on two fundamental notions from evolutionary biology: orthology and paralogy, which are the two fundamentally different types of homologous relationships between genes (Sonnhammer and Koonin, 2002).

A clear distinction between orthologs and paralogs is critical for the construction of a robust evolutionary classification of genes and reliable functional annotation of newly sequenced genomes. Orthologous genes derive from a single ancestral gene in the last common ancestor of the compared species, and paralogous genes are related via gene duplication (Koonin, 2005). In this context, ortholog prediction is often used in genome annotation, gene function characterization, evolutionary genomics, and identification of conserved regulatory elements. Errors in ortholog prediction can affect such studies and associated downstream analyses (including functional genomics and proteomics analyses), so

there is an increasing interest in high quality ortholog prediction (Fulton et al., 2006).

In general, several methods used to infer orthologs are based on reciprocal best BLAST hits relationships (bet), using all-against-all comparison. The COG (Clusters of Orthologous Groups—including prokaryotes and unicellular eukaryotes) and KOG (Clusters of Eukaryotic Orthologous Groups) databases were built using such methods, after masking low-complexity and predicted coiled-coil regions (Tatusov et al., 2003). In COG, an orthologous set is based on the notion that any group of at least three proteins from distant genomes is more similar to each other than they are to any other protein from the same genome (Wilson Kreychman and Gerstein, 2000).

However, the accuracy of the methods for orthology inference depends on the evolutionary distance. The greater the distance, the less accurate the inference (Reeves et al., 2009; Wilson Kreychman and Gerstein, 2000). To overcome this limitation, powerful sequence comparison methods have been developed, such as Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) (Altschul et al., 1997) and Hidden Markov Models (HMM) (Eddy, 2006), which use sequence profiles built up of groups of related sequences to identify remote homologue proteins (Reeves et al., 2009). An example of a tool that uses PSI-BLAST is ESG (Chitale et al., 2009), which identifies orthologs and performs functional annotation with higher sensitivity than traditional methods based on BLAST.

The aim of this study is to use an orthology-based approach to: (i) infer the proteomic core between five protozoan species, extending the original analysis carried out by El-Sayed et al., (2005) by adding other two pathogenic protozoa from different groups, and (ii) use these groups from proteomic to phylogenomic analysis.

While the blast-based OrthoMCL software (Li et al., 2003) was used to infer orthologs among protozoa species, OrthoSearch (Cruz et al., 2010) was used to infer Protozoa-Unicellular and Protozoa-Eukaryote orthologs. The latter software was originally designed as a scientific workflow for orthology inference using an HMM-based approach. The InterPro package (Mulder et al., 2005) was used for the *in silico* validation of annotations of the identified orthologs. Complete proteomes of five protozoan species (*T. brucei*, *T. cruzi*, *L. major*, *E. histolytica*, and *P. falciparum*) were analyzed. The proteome sequences of these protozoa and *in silico* experimental results (including orthology data) are available in public databases such as ProtozoaDB (Dávila et al., 2008; BiowebDB, 2009) (<http://protozoadb.biowebdb.org>).

Material and Methods

Figure 1 illustrates the methodology used in this work and the following sections provide details about the methodology.

Dataset

The protein dataset of five protozoa (*T. cruzi*, *T. brucei*, *L. major*, *E. histolytica*, and *P. falciparum*), available in Fasta format, was obtained from ProtozoaDB version 1 (Dávila et al., 2008). We used the CD-HIT program (Li, 2006) with a cutoff of 100% of identity to remove the redundancy of the sequences. The orthologous groups from COG and KOG

FIG. 1. Fluxogram illustrating the methodology used in this study.

were downloaded from NCBI (<http://www.ncbi.nlm.nih.gov/COG/>) in August 2008.

Orthology identification

OrthoSearch—COG/KOG against Protozoa. OrthoSearch (Cruz et al., 2010) is a scientific workflow developed under VisTrails Scientific Workflow Management System (Callahan et al., 2006). It was used to infer orthologous groups between protozoa and COG/KOG. The COG and KOG orthologous groups were used as input to generate Hidden Markov Model profiles (HMMp), and then those HMMp were searched against all protein sequences from the five protozoan genomes. Briefly, OrthoSearch workflow uses MAFFT software (Katoh et al., 2005) with default parameters to align each original COG or KOG group present in our dataset. Then, each multiple alignment is used to build and calibrate the HMMp using the HMMER software package (Eddy, 2006). Two separate HMMp datasets were obtained, the first, from COG multiple alignments, is called COG-HMMp, and the second one, from KOG multiple alignments, is named KOG-HMMp. Both HMMp datasets were searched

against all protozoan proteins available for each species, using the *hmmsearch* (HMMp against protozoan proteins) and *hmmpfam* (protozoan proteins against HMMp) programs of HMMER with cutoff *e*-value 0.1. The reciprocal best hit found between HMMp and protozoan proteins represent orthologs, according to the original description of OrthoSearch.

The functional overlap between the groups of the COG and KOG were verified using a table kindly provided by Dr. Michael Y. Galperin (personal communication), and comparing the accession numbers of protozoan proteins in the groups detected by NCBI's COG and KOG.

OrthoSearch—Protozoa orthologous groups against Protozoa. To increase the sensitivity of distant homologs detection inside the five protozoan genomes, a second run of OrthoSearch was executed using only protozoan proteomes as a target. Briefly, the Protozoa-COG/KOG orthologous groups inferred by the first run of OrthoSearch were aligned (Brayton et al., 2007) and new HMMp (Protozoa-COG/KOG-HMMp) were built (Chitale et al., 2009). These Protozoa-COG/KOG-HMMp were used to search for additional distant

homologs among the five protozoan proteomes (again using the OrthoSearch workflow).

OrthoMCL. The protein sequences of the five protozoa in our dataset were used to perform ortholog identification with OrthoMCL (Li et al., 2003). All proteins in the five protozoa were compared against themselves in a search for homologous proteins. In this analysis, the cutoff *P*-value used was $1e^{-5}$ as proposed by Li et al. (2003). All results were loaded into ProtozoaDB using a GUS (Davidson et al., 2001) plugin specific for that.

Comparison between OrthoMCL and OrthoSearch. All orthologous groups common to the five protozoa inferred by OrthoSearch and OrthoMCL were compared and evaluated. Each protein of the orthologous groups inferred by OrthoSearch was compared with the proteins of the orthologous groups inferred by OrthoMCL (through its accession number). The groups were considered either (i) equal when the same proteins of the five protozoa identified by OrthoSearch were present in the OrthoMCL orthologous groups or (ii) different when no protein was found in common.

Re-annotation

The Protozoa-COG/KOG orthologs identified by OrthoSearch were: (i) carefully analyzed aiming to use them to re-annotate proteins from the five protozoa; and (ii) loaded into ProtozoaDB (Dávila et al., 2008) using another scientific workflow called OrthoLoad. The re-annotation process mentioned in item (i) was manually performed. For each “Protozoa-COG/KOG” orthologous group, the original COG/KOG description was confronted with the description available at NCBI for the protozoan protein. If the annotation was the same (at the family level), then we considered the annotation of the protozoan protein as “confirmed.” If the annotation of the protozoan protein was different from the COG/KOG description, then the COG/KOG annotation was transferred and the protozoan protein considered as “re-annotated.” If the protozoan protein had no annotation or its description was “unknown” or “hypothetical,” then we transferred the COG/KOG annotation and considered the protozoan protein “annotated”.

Comparing re-annotation of hypothetical and unknown proteins with TritypDB

To validate our semi-automatic re-annotation of hypothetical proteins (and unknown function proteins) in the Trityp genomes, the GenBank GI of those re-annotated proteins was mapped (April, 2012) into TritypDB 4.0 (<http://tritypdb.org/tritypdb/>) (Aslett et al., 2010) and the annotations (OrthoSearch semi-automatic annotation versus TritypDB curated annotation) manually compared.

Re-annotation validation using InterPro

The annotation and validation module of OrthoSearch uses the fastacmd program from the BLAST package (Altschul et al., 1990) to obtain FastA formatted protein sequences that correspond to the reciprocal best hits or inferred orthologs, and then submit them to InterPro (Mulder et al., 2005) analysis. This module was used to validate the re-annotation

performed with OrthoSearch and COG-HMMp or KOG-HMMp in protozoan proteomes. Briefly, the protozoan proteins of each Protozoa-COG/KOG orthologous groups were analyzed by InterPro 1.4 (interproscan program using all its databases) on local Xeon quad-core servers. If the annotations transferred from COG/KOG to protozoan proteins matched with at least one of the results obtained by InterPro, then we considered the OrthoSearch annotation as “validated,” if not, the annotation was considered “non-validated.”

Phylogenomic analysis

In order to evaluate if the shared orthologs between (i) the five protozoa and KOG groups (only those orthologs where the five protozoan proteins were not found in other groups inferred by COG-pHMM—eukaryote exclusives); and (ii) the five protozoa, COG groups, and KOG groups (only those orthologs where the same five proteins from protozoa were detected by COG-pHMM and KOG-pHMM) are reliable markers for species tree reconstruction, a supertree-based approach was tested and the resulting trees compared to the protozoa species tree obtained by Ocaña and Dávila (2011). Briefly, phylogenetic trees were inferred using the ARPA pipeline, developed as part of a protozoan species tree study (Ocaña and Dávila, 2011). The multifasta files of each of the two datasets (i and ii) were aligned using MAFFT (Katoh et al., 2005) with default parameters. These alignments were used to construct NJ trees with the PHYLIP 3.69 package and 1,000 bootstraps replicates. The phylogenetic trees were manually concatenated, and the program Clann (Creevey and McInerney, 2005) was used to construct super-trees (one for each of the two datasets).

Results

Dataset

An initial dataset of 57,512 protein sequences (without redundancy) was obtained. Non-redundant sequences were distributed as follows: 20,322 of *T. cruzi*, 9,639 of *T. brucei*, 8,189 of *L. major*, 11,116 of *P. falciparum*, and 8,246 of *E. histolytica*. A total of 2855 orthologous groups from COG and 3578 from KOG were downloaded from NCBI (only characterized groups).

Orthology identification

OrthoSearch—COG/KOG against Protozoa. Using the COG-HMMp dataset (2855 HMM profiles), a total of 3901 orthologs were identified, whereas using the KOG-HMMp (3578 HMM profiles) a total of 7473 were found. Table 1 shows the distribution of these orthologous clusters in the five protozoan species and the estimated total coverage of the proteome.

The orthologous groups of protozoa inferred by OrthoSearch are distributed in most functional categories of the COG/KOG. Figure 2 shows the overall distribution between the functional categories of COG and KOG found in protozoa. Using COG-HMMp, 753 groups were obtained among the two species of *Trypanosoma*, 1.73% (13/753) of these groups were found only in these parasites. Using KOG-HMMp, the proteomic core of *Trypanosoma* was 1497 groups, where 2.4% (36/1,497) groups are present only in this genus. Extending the analysis to the proteomic core of

TABLE 1. DISTRIBUTION OF ORTHOLOGOUS GENES OF KOG AND COG FOR EACH SPECIES OF PROTOZOA

Organism	Total hits with COG	Total hits with KOG	Total coverage of genomes
<i>E. histolytica</i>	672	1261	8.15% (COG) 15.30% (KOG)
<i>P. falciparum</i>	724	1361	6.51% (COG) 12.24% (KOG)
<i>L. major</i>	863	1615	10.54% (COG) 19.72% (KOG)
<i>T. brucei</i>	798	1587	8.28% (COG) 16.47% (KOG)
<i>T. cruzi</i>	844	1649	4.15% (COG) 8.11% (KOG)

Trityp, 719 groups were inferred using COG-HMMp, where 103 were found only in these three protozoa, and 1405 groups were inferred using KOG-HMMp, with 245 present only in these three species. A total of 418 and 704 orthologous groups were found in the five species, using COG-HMMp and KOG-HMMp, respectively, representing a proteomic core shared by all these five parasites. Figures 3 and 4 show Venn diagrams with the number of groups formed considering the protozoa analyzed with COG-HMMp and KOG-HMMp. A total of 158 of these groups are redundant (the same five proteins from protozoa were detected by COG-HMMp and KOG-HMMp), representing orthologs shared among unicellular species and eukaryotes. Both COG-HMMp and KOG-HMMp were able to detect 63 and 328 exclusives orthologous groups, respectively (no other group with the same five proteins was detected).

OrthoSearch—Protozoa orthologous groups against Protozoa. After running the OrthoSearch workflow (for a second time) using HMMp built by groups formed by proteins from two, three, or four protozoa (inferred on the first run by COG-HMMp and KOG-HMMp), further 131 groups were detected by groups inferred by COG-HMMp and 229 groups detected by groups inferred by KOG-HMMp, resulting in a total of 549 (using COG-HMMp inferred groups) and 933 (using KOG-HMMp inferred groups) orthologous groups for the five protozoa after the second run.

OrthoMCL

Using OrthoMCL with its default “all-against-all” search on the five protozoa, a total of 7072 orthologous groups and 1865 paralogous groups were found. A total of 498 orthologous groups are shared by the five species. The proteomic core shared between the two species of trypanosomes was inferred, with 6433 orthologous groups, including 12.23% (787/6433) unique to this genus. The proteomic core of Trityp is slightly small, with 5608 formed groups, of which 76.27% (4277/5608) are exclusive. Figure 5 shows a Venn diagram with the number of groups formed considering the protozoa studied with OrthoMCL.

Comparison between OrthoMCL and OrthoSearch

OrthoMCL detected 498 groups of orthologs shared by the five protozoa, and OrthoSearch detected 1122 orthologous groups shared by the five protozoa (704 by KOG-HMMp and 418 by COG-HMMp). From the orthologous groups inferred by OrthoSearch, 38.15% (428/1122) were also found by OrthoMCL (131 overlapping with groups inferred by COG-HMMp and 297 by KOG-HMMp) and 508 groups (187 inferred by COG-HMMp and 321 by KOG-HMMp) were exclusively found by OrthoSearch. On the other hand,

FIG. 2. Distribution of OrthoSearch hits in the functional categories of COG proteins from protozoa detected by OrthoSearch (using COG and KOG groups), classified using the functional categories of COG (<http://www.ncbi.nlm.nih.gov/COG/grace/fiew.cgi>).

FIG. 3. Distribution of orthologous groups inferred by OrthoSearch (using COGs) between the five species of protozoa. The Venn diagram shows the distribution of the orthologous groups from the five protozoa inferred by OrthoSearch using the COG dataset. The green area indicates the genomic core of the five protozoa. The colored areas marked with the character * indicate the genomic core of the Tritryp. It summarizes 719 groups (59 + 103 + 139 + 418). The blue area indicates the COG groups of each single protozoan. The character § indicates the number of exclusive groups for the genomic core of the five protozoa.

18.47% (92/498) groups shared by the five protozoa were detected exclusively by OrthoMCL. Table 2 shows the results of such analysis.

Re-annotation

To verify the effectiveness of our re-annotation methodology using OrthoSearch, the annotation transferred to protozoa (with COG-HMMp and KOG-HMMp) was manually compared to the previously existing NCBI annotation. Figure 6 describes groups inferred by COG-HMMp, and Figure 7 describes groups inferred by KOG-HMMp. Figure 6 shows three kinds of possible results of annotation condition, and their values are represented as percentages: (i) the previous annotations assigned by NCBI agree with the obtained by our method (that confirms the NCBI annotation); (ii) the annotations assigned by OrthoSearch disagree with previous annotation provided at NCBI (annotation obtained was different from original); and finally, (iii) the category of proteins that do not have annotation assigned at NCBI (hypothetical), although OrthoSearch was still able to infer a putative function for them.

Among the total of proteins that had their annotation manually verified by us, it was observed that: (i) 79.03% (3083/3901) and 71.99% (5380/7473) agree with previous annotations assigned at NCBI; (ii) 0.72% (28/3901) and 0.52% (39/7473) disagree; (iii) 20.25% (790/3901) and 27.49% (2054/7473) were annotated as hypothetical or do not have annotations assigned at NCBI, using COG-HMMp and KOG-HMMp, respectively.

Comparing the OrthoSearch re-annotation of hypothetical and unknown proteins with TritrypDB

Using OrthoSearch with COG-HMMp and KOG-HMMp, 532 and 1383 proteins (previously described as hypotheticals at NCBI) were re-annotated in the Tritryp proteome, respectively. The NCBI hypothetical and unknown proteins could be classified into three groups: (i) sequences also annotated as hypothetical or unknown in TritrypDB (Aslett et al., 2010); (ii) sequences not present in TritrypDB; and (iii) sequences with functional annotation in TritrypDB. This exercise provided the following data: (i) 90.60% of NCBI hypothetical sequences (482/532 of the re-annotations obtained with COG-HMMp) and 90.02% (1245/1383 of the re-annotations obtained with KOG-HMMp) remain uncharacterized in TritrypDB; (ii) 2.25% (12/532 of the re-annotations obtained with COG-HMMp) and 2.53% (35/1383 of the re-annotations obtained with KOG-HMMp) are not present in TritrypDB; and (iii) only 7.14% (38/532) (COG-HMMp) and 7.44% (103/1383) (KOG-HMMp) have functional annotation at TritrypDB. Table 3 shows the results of these analyses.

Re-annotation validation using InterPro

We have also compared annotations inferred by InterPro with annotations inferred by OrthoSearch, showing that the relationship of reciprocal best hit with OrthoSearch is effective. The results of InterPro in 94.27% (10,723/11,374) of cases corroborate the results obtained by OrthoSearch, and in 4.58% (521/11,374) of InterPro results, the hits were against

FIG. 4. Distribution of orthologous groups inferred by OrthoSearch (using KOGs) between the five species of protozoa. The Venn diagram shows the distribution of the orthologous groups from the five protozoa inferred by OrthoSearch, using the KOG dataset. The *green area* indicates the genomic core of the five protozoa. The *colored areas* marked with the character * indicate the genomic core of the Tritryps; it summarizes 1405 groups ($704 + 284 + 245 + 172$). The *blue area* indicates the KOG groups of each single protozoan. The character § indicates the number of exclusive groups for the genomic core of the five protozoa.

sequences without description or hypotheticals. Table 4 shows the results of such analysis.

Phylogenomic analysis

The orthologous groups shared by the five protozoans and the orthologous groups from NCBI were used in phylogenomic analysis. Two datasets were generated: (i) 158 groups shared by the five protozoa, the COG (unicellular orthologous groups) and KOG groups; and (ii) 328 shared only by the five protozoa and KOG groups. For each dataset, NJ trees were inferred (one for each group) and concatenated for super-tree analysis. Figures 8 and 9 show the result of analyses using groups from dataset (i) and (ii), respectively. Four well-supported subclades were inferred among Eukaryotes (Fig. 8) with Protozoa forming a separate clade.

Discussion

Most of the drugs used in the treatment and prevention of NTDs are extremely toxic and debilitating, such as nifurtimox for Chagas disease (suspended because it causes many side effects) and the only currently available drug in the Brazilian market called benznidazole (with low efficacy in the chronic phase of the Chagas disease). Furthermore, drug resistance has been observed in major public health impact studies of Chagas disease, malaria, and leishmaniasis (Lindoso and Lindoso, 2009).

Genomic and proteomic studies of these protozoa can help in the identification of stage-specific pathways, not documented, that may be suitable for the development of new drugs targets (Atwood et al., 2005). Comparing genomes of different pathogens, for example, has helped to find

parison of the genomes of protozoa belonging to different taxonomic groups can lead to the development of drugs that have a broad spectrum, and potentially less toxic to mammalian hosts. As a consequence, it increases the number of drugs available for treating these diseases, since, during the last 20 years, few drugs have been developed to treat diseases caused by these protozoa (Trouillier et al., 2002).

Dataset

The exclusion of poorly characterized COG/KOG groups (categories R: general function prediction only, and S: function unknown) was done because the main objective is to perform functional re-annotation and these groups have no functional description. Moreover, the redundancy removal of sequences of protozoa using CD-HIT program is desirable because our initial proteomic dataset is hybrid (it contains sequences from GenBank and RefSeq), and these datasets are known to overlap (RefSeq generated from Genbank).

FIG. 6. Pie charts representing the annotation transfer efficiency obtained with OrthoSearch. The figure shows the percentages of annotations of protozoan proteins generated by OrthoSearch (using COG). The annotations are distributed according to three categories: (i) annotations that agree with previous proteins annotations are *blue*, (ii) the hypothetical or non-annotated proteins are *red*, and (iii) the annotations that disagree or are not confirmed with previous annotations are *green*.

Orthology identification

OrthoSearch. With the aid of OrthoSearch workflow, it was possible to infer the orthology of protozoan proteins with values ranging from 4.15% (*T. cruzi*/COG) to 19.72% (*P. falciparum*/KOG). While this percentage may seem low, the percentage values are encouraging if one considers that the utilization of the seed orthologous datasets (COG/KOG) was built from a limited number of sequenced organisms (66 prokaryotes for COG and only seven eukaryotes for KOG), and also considering that these organisms are not taxonomically close to Protozoa. In other words, if one takes into account that the total number of COG/KOG orthologous groups available at NCBI database is 4872 and 4852, respectively, and the reduced amount of COG/KOG groups (2855 orthologous groups from COG and 3578 from KOG) used by OrthoSearch, then our methodology worked well. Extrapolating results, the KEGG orthologous groups (KO) could be used in future experiments to seed new HMM-profiles (to search against Protozoa dataset) and hopefully provide even better results because KO has a larger number of genomes (1483 prokaryotes and 151 eukaryotes, including five trypanosomatids and other protozoans).

Besides that, if one compares such numbers with the whole genome of a trypanosomatid (ranging from 8,311 to 12,000

proteins), it is possible to verify that the coverage of the COG/KOG orthologous groups compared to the genome of a trypanosomatid is also too short. The trypanosomatids genomes have a high degree of duplication (paralog families with many members) (Berriman et al., 2005; El-Sayed et al., 2005a), and OrthoSearch detects only orthologous groups presenting only one member per family (reciprocal best hit). A higher number of orthologous groups could be potentially detected in the protozoan genomes if a larger number of seed orthologous groups could be used instead of COG/KOG as part of the OrthoSearch workflow. This might be the case of KO, having about 14,000 orthologous groups, covering almost all proteins that exist in 1634 genomes (Ogata et al., 1999).

The orthologous groups detected by OrthoSearch in the five protozoa are distributed over all functional categories of the COG/KOG used in this analysis, and many orthologs belong to category J (translation, ribosomal structure, and biogenesis), 813 hits (20.84% of total hits) obtained by COG-HMMp and 1057 hits (14.14% of total hits) by KOG-HMMp. Such result was expected because this category has most of the universal orthologous groups, as they are inferred from existing genes in all species according to Ciccarelli et al. (2006). All of these 31 universal groups were detected in the five protozoans. The COG/KOG functional categories L, O, and A have a large number of orthologs (L: 298/430, O: 368/1070

FIG. 7. Pie charts representing the efficiency (percentages) of annotations of protozoan proteins generated by OrthoSearch (using KOG). The annotations are distributed according to three categories: (i) annotations that agree with previous proteins annotations are *blue*, (ii) the hypothetical or non-annotated proteins are *red*, and (iii) the annotations that disagree or are not confirmed with previous annotations are *green*.

and A: 70/663) detected by COG and KOG, respectively, and are also related to translation, transcription, and repair machinery.

Using the KOG-HMMp, we have obtained the highest number of groups shared by the five organisms (704). This

TABLE 3. COMPARISON BETWEEN ANNOTATIONS OBTAINED BY ORTHOSEARCH WITH TRITRYPDB

<i>COG</i>			
<i>Organism</i>	<i>Hypothetical proteins annotated by OrthoSearch</i>	<i>Hypothetical proteins annotated by OrthoSearch with annotation in TritypDB</i>	<i>Do not exist in TritypDB</i>
<i>T. cruzi</i>	179	10	6
<i>T. brucei</i>	155	17	3
<i>L. major</i>	198	11	3
Total	532	38	12
<i>KOG</i>			
<i>T. cruzi</i>	488	18	17
<i>T. brucei</i>	431	52	17
<i>L. major</i>	464	33	1
Total	1,383	103	35

result was expected because these are groups formed by sequences of eukaryotes while the COG was built using mostly prokaryotes. However, it was possible to obtain a considerable number of hits using COG-HMMp: 418 orthologs shared by the five genomes, (~59% of the number obtained by KOG-HMMp) showing that it is interesting to use both COG-HMMp and KOG-HMMp. Furthermore, as the correspondence mappings between COG and KOG were kindly provided by Dr. Michael Y. Galperin (personal communication), it was possible to identify COG with no homologs in eukaryotes, and KOG with no homologs in COG. From the 418 protozoan orthologous groups inferred using COG-HMMp, 24.40% (102/418) have no relationship with any KOG group. When analyzing the correspondence mapping between COG and KOG, it was also observed that 10.52% (44/418) of the COG-HMMp that found orthologs in protozoa have homologs in KOG; however, those corresponding 44 KOG-HMMp were not able to detect orthologs in protozoa. Based on the correspondence mappings, those 44 proteins in the five protozoa are shared by prokaryotes and eukaryotes, but based on the impossibility of OrthoSearch to detect those 44 KOG-HMMp in protozoa, then it is also possible that those proteins in protozoa are more related to COG organisms than KOG. Given the sensitivity of our HMM-based tool, a possible explanation is that those 44 orthologs, not detected by KOG-HMMp in protozoa, are so

TABLE 4. RE-ANNOTATION VALIDATE USING INTERPRO (INTERPROSCAN)

	Confirmed by InterPro	Not confirmed by InterPro	Hypothetical by InterPro	Without results by InterPro
COG	3,716	49	12	124
KOG	7,007	81	0	385

divergent that OrthoSearch was not able to detect them. The other 272 groups (of 418 orthologous groups inferred by COG-HMMp for the five parasites) were also inferred by KOG with the same functional relationship with COG.

When comparing the groups that are shared by the five protozoa obtained with the COG-HMMp and KOG-HMMp (Fig. 2), it was observed that two categories with the greatest number

of hits were the same for both, J and O (post-translational modification, protein turnover, chaperones) and the main difference between the categories is that with the COG-HMMp the third largest category was the L (replication, recombination, and repair), while with the KOGs-HMMp it was category A (RNA processing and modification).

OrthoMCL. The OrthoMCL program inferred a large number of groups of inparalogs proteins: 1865 genes for the five organisms. *E. histolytica* has 301 proteins with 763 copies (mean of 2.53 genes/family), *L. major* has 82 proteins with 254 copies (3.1 genes/family), *P. falciparum* has 401 proteins with 6171 copies (with the mean of 15.39 proteins/family), *T. brucei* has 139 proteins with 918 copies (6.60 proteins/family), and finally *T. cruzi* has 5469 copies in 942 proteins (5.81 proteins/family). This large number of duplications was expected because some of these organisms have a high gene duplication rate (Li et al., 2003).

FIG. 8. COG-KOG-Protozoa Super-tree of concatenated NJ trees from orthologous groups, obtained using the Clann software, with default parameters and bootstrap analysis (100 replicates). Ath, *Arabidopsis thaliana*; Cel, *Caenorhabditis elegans*; Dme, *Drosophila melanogaster*; Ecu, *Encephalitozoon cuniculi*; ehistolyti, *Entamoeba histolytica*; Hsa, *Homo sapiens*; Imajor, *Leishmania major*; pfalciparu, *Plasmodium falciparum*; Sce, *Saccharomyces cerevisiae*; Spo, *Schizosaccharomyces pombe*; tbrucei, *Trypanosoma brucei*; tcruzi, *Trypanosoma cruzi*; Organisms from COG: <ftp://ftp.ncbi.nih.gov/pub/COG/COG/org.txt>.

As in OrthoSearch, it was possible to identify with OrthoMCL a proteomic core among the five organisms, as 498 groups are shared by all five organisms. The number of orthologous groups increases considerably when considering only the genomes of *Trityp*, sharing 5608 groups; this was

Although the total number of detected orthologous groups in OrthoMCL is larger than OrthoSearch, the OrthoSearch approach is more sensitive in detecting distant homologies. The OrthoMCL can infer 6433 groups between two

organisms of the same genus (*Trypanosoma*), against 753 and 1497 (COG and KOG, respectively) inferred by OrthoSearch. This represents an increase of 854.31% (OrthoMCL vs. OrthoSearch-COG-HMMP) and 429.72% (OrthoMCL vs. OrthoSearch-KOG-HMMP).

On the other hand, the number of shared groups to the five organisms was ~125% higher using OrthoSearch than OrthoMCL. For example, a total of 1122 groups (704 with KOG-HMMP and 418 with COG-HMMP) against 498 using OrthoMCL were inferred. This fact demonstrates that the methodology used by OrthoSearch is more sensitive and effective when it is necessary to analyze the genomes of not too close taxonomically related organisms. This is due to the fact that OrthoSearch uses a profile-based approach, while OrthoMCL uses the BLAST program, which is less sensitive than techniques based on profiles (Eddy, 2006). When analyzing the proteomic core inferred by the two methods, we have observed an overlap of 428 groups (most of 498 OrthoMCL groups were also inferred by OrthoSearch), while OrthoSearch was able to infer 508 groups that were not inferred by OrthoMCL. The groups that were found exclusively by OrthoMCL are mostly hypothetical proteins (47.8% of 92 groups). This suggests that these orthologous groups might be exclusive of Protozoa, since OrthoSearch did not infer orthologs to COG/KOG. However, this has to be carefully considered as both COG/KOG have not been publicly updated recently and the addition of new eukaryote genomes to them could occasionally help identifying orthologs with Protozoa.

Re-annotation

Through the comparison between the pre-existing annotations (at NCBI) and the ones inferred by OrthoSearch we can see that, in the case of previously well annotated genes, OrthoSearch was able to confirm this annotation.

In the few cases where there were mismatches, it was due to poorly annotated proteins (less than 1% of the cases) (Fig. 6). When using COG-HMMP, 783 hypothetical proteins were re-annotated in five organisms, and when using KOG-HMMP 2,046 also in the same five organisms, representing 2.58% and 6.76% of total hypothetical in the five proteomes, respectively. This demonstrates that our methodology is also useful after manual annotation, especially for proteins previously considered hypothetical, reducing the problems of the “known unknowns.” For instance, Galperin and Koonin (2004) highlighted that in any newly sequenced bacterial genome, as many as 30%–40% of the genes do not have an assigned function. This figure is even higher for archaeal and eukaryotic genomes (Carlton et al., 2008). By using OrthoSearch we were able to re-annotate ~11,000 proteins from a total of ~57,000.

Comparing OrthoSearch re-annotation of hypothetical and unknown proteins with TritypDB

Our methodology was able to infer function for more than a thousand proteins annotated as hypothetical or unknown function in the curated TritypDB. This represents approximately 90% (482/532 using COG-HMMP and 1245/1383 using KOG-HMMP) of the hypothetical proteins, which OrthoSearch was able to re-annotate for the three proteomes (Trityp).

Since TritypDB is the main repository for trypanosomatids curated sequences, we assume no other *in silico* approach than ours was able to re-annotate these protozoan hypothetical proteins.

These findings support the usefulness of the HMM-based methodology to infer function for previously uncharacterized proteins in different genomes.

Re-annotation validation using InterPro

The results of InterPro (interproscan) confirm, in most cases, the annotations obtained with HMMER package. There are only few cases where the annotation provided by InterPro, for a given sequence, did not match with HMMER. Interproscan uses a series of databases and similarity search tools, in addition to other features, to infer the sequence annotation and it was useful in validating the re-annotation performed with OrthoSearch. This shows that it is possible to use OrthoSearch in automatic re-annotation with a high degree of accuracy. However, InterPro does not have orthologous groups of KOG and COG in its database, so it only serves as a complementary method in our analyses.

Phylogenomics analysis

Using a super-tree approach, it was possible to test the orthologous groups shared by the five protozoa. Two datasets were used, as described in the Material and Methods section, and the resulting trees were carefully analyzed. Using the groups shared by COG, KOG, and Protozoa, it was possible to obtain a good resolution on species tree (Fig. 8). The three domains of life (Archaea, Bacteria, and Eukarya) were separated with good support values (bootstrap analysis), and the Archaea group is shown as being closer to Eukarya than Bacteria, as expected, and previously shown on the tree of life (Ciccarelli et al., 2006). The Bacteria groups are properly separated as well. The phylogenomic analysis showed four well-supported clades for Eukarya, discriminating Multicellular [(i) human, fly, plant and worm] and Unicellular [(ii) Yeast, (iii) Fungi, and (iv) Protozoa] species. The species tree in Figure 8 corroborates classical and Tree of Life findings, but also shows the incompleteness of KOG by missing Protozoan and other Eukaryote species. This has also an impact in our study, corroborating that many genes in protozoa cannot be found just using the orthologs available in KOG, consequently more genes and orthologs could be identified in protozoa (and be potentially re-annotated) by using a combination of KOG, COG, and protozoan-specific orthologs. In fact, our experiment, using a second run of OrthoSearch, using protozoan orthologs, resulted in an increase of protozoan orthologs detection up to 549 (131 more than the initial findings using COG-HMMP inferred groups) and 933 (229 more than the initial findings using KOG-HMMP groups).

Conclusions

This study shows that the HMM-based methodology developed by us is effective for the re-annotation of poorly annotated sequences or hypothetical proteins. It can be used in addition to traditional methodologies like OrthoMCL. We have observed that 94.27% (10,723/11,374) of the total number of proteins that were accurately re-annotated by OrthoSearch (either using COG-HMMP or KOG-HMMP),

had their annotation confirmed by InterPro and also cross-validated using TritypDB. With the aid of OrthoSearch, it was also possible to infer the core proteome of the five studied organisms with greater sensitivity, showing that our profile-based approach is better for such purpose. Considering only the non-annotated proteins, that are annotated as hypothetical or unknown, our profile-based methodology was able to annotate between 2.58% (782/30,241) and 6.76% (2,044/30,241), using COG-HMMp and KOG-HMMp, respectively. By using OrthoSearch we achieved genome re-annotation coverage between 4.15% to 10.54% using COG-HMMp and 8.11% to 19.72% when using KOG-HMMp. The phylogenomic analysis of this work showed that it is possible to use the orthologous groups inferred by OrthoSearch in order to infer the phylogenetic relationship among Protozoa and the organisms from COG and KOG database. These results encourage us to develop further investigations on OrthoSearch with HMMER version 3.0 and using profiles based on a more comprehensive orthologous database such as KEGG Orthology (KO).

Finally, it is noteworthy that the approach presented herein may also lend itself for applications in global health, for example, in the case of novel drug target discovery against pathogenic organisms previously considered difficult to research with traditional drug discovery tools.

Acknowledgments

We thank for the financial support received from the Brazilian National Research Agencies CAPES (AMRD) and CNPq (MM) and Rio de Janeiro State Research Agencies (SMSC, MLMC, MM) that enabled this research, and wish to express our gratitude to Dr. Michael Y. Galperin (NCBI) for valuable criticism on an earlier draft version of the manuscript. Thanks also to five anonymous referees and Prof. Vural Özdemir for valuable criticism and suggestions.

Author Disclosure Statement

The authors declare that they have no competing interests or financial disclosures. AMRD is the coordinator of the BioWebDB Consortium funded by CAPES and CNPq. SMSC, MLMC and MM are also members of the BioWebDB Consortium.

References

Altschul SF, Madden T, Schaffer AA, et al. (1997). BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25, 3389–3402.

Altschul SF, Gish W, and Miller W. (1990). Basic local alignment search tool. *J Mol Biol* 215, 403–410.

Aslett M, Aurrecochea C, Berriman M, et al. (2010). Tri-TrypDB: A functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res* 38, D457–D462.

Atwood JA, Weatherly DB, Minning TA, et al. (2005). The *Trypanosomacruzi* proteome. *Science* 309, 473–476.

Batista M, Marchini F, Celedon P, et al. (2010). A high-throughput cloning system for reverse genetics in *Trypanosoma cruzi*. *BMC Microbiol* 10, 259.

Berriman M, Ghedin E, Hertz-Fowler C, et al. (2005). The genome of the African trypanosome *Trypanosoma brucei*. *Science* 309, 416–422.

BioWebDB Consortium—Comparative genomics approaches [http://biowebdb.org/]. Last access: April, 2012.

Brayton KA, Lao AO, Herndon DR, Hannick L, and Kappmeyer LS. (2007). Genome sequence of *Babesia bovis* and comparative analysis of Apicomplexan Hemoprotozoa. *PLoS Pathogens* 3, 1401–1413.

Brehelin L, Florent I, Gascuel O, and Marechal E. (2010). Assessing functional annotation transfers with inter-species conserved coexpression: Application to *Plasmodium falciparum*. *BMC Genomics* 11, 35.

Callahan SP, Freire J, Santos E, et al. (2006). VisTrails: Visualization meets data management. Proceedings of the 2006 ACM SIGMOD international conference on Management of data. Chicago, USA, 745–747.

Carlton JM, Adams JH, Silva JC, et al. (2008). Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* 455, 757–763.

Carlton J, Silva J, and Hall N. (2005). The genome of model malaria parasites, and comparative genomics. *Curr Issues Mol Biol* 7, 23–38.

Cavalier-Smith T. (2010). Deep phylogeny, ancestral groups and the four ages of life. *Philosoph Trans Royal Soc B: Biol Sci* 365, 111–132.

Ciccarelli FD, Doerks T, von Mering C, et al. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* 311, 1283–1287.

Chitale M, Hawkins T, Park C, and Kihara D. (2009). ESG: Extended similarity group method for automated protein function prediction. *Bioinformatics* 25, 1739–1745.

Collier P. (2007). *The Bottom Billion: Why the Poorest Countries are Failing and What Can Be Done About It*. 1st ed. Oxford University Press.

Creevey CJ, and McInerney JO. (2005). Clann: Investigating phylogenetic information through supertree analyses. *Bioinformatics* 21, 390–392.

Cruz SMS, Batista V, Silva E, et al. (2010). Detecting distant homologies on protozoans metabolic pathways using scientific workflows. *Intl J Data Mining Bioinform* 4, 256–280.

Davidson S, Crabtree J, Brunk BP, et al. (2001). K2/Klesli and GUS: Experiments in integrated access to genomic data sources. *IBM Systems J* 40, 521–531.

Dávila AMR, Mendes P, Wagner G, et al. (2008). ProtozoaDB: Dynamic visualization and exploration of protozoan genomes. *Nucleic Acids Res* 36, D547–D552.

Doolittle WF. (1998). You are what you eat: A gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genetics* 8, 307–311.

Eddy SR. (1996). Hidden Markov models. *Curr Opin Struct Biol* 6, 361–365.

El-Sayed NM, Myler PJ, Bartholomeu DC, et al. (2005a). The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* 309, 409–415.

El-Sayed NM, Myler PJ, Blandin G, et al. (2005b). Comparative genomics of Trypanosomatid parasitic protozoa. *Science* 309, 404–409.

Feasey N, Wansbrough-Jones M, Mabey DCW, and Solomon AW. (2010). Neglected tropical diseases. *Br Med Bull* 93, 179–200.

Fitch W. (1970). Distinguishing homologous from analogous proteins. *Systematic Biol* 19, 99–113.

Fulton DL, Li YY, Laird MR, Horsman B, and Brinkman F. (2006). Improving the specificity of high-throughput ortholog prediction. *BMC Bioinform* 7, 220.

- Galperin MY, and Koonin EV. (2004). Conserved hypothetical proteins: Prioritization of targets for experimental study. *Nucleic Acids Res* 32, 5452–5463.
- Hall N, and Carlton J. (2005). Comparative genomics of malaria parasites. *Curr Opin Genetics Devel* 15, 609–613.
- Hastings IM, Watkins WM, and White NJ. (2002). The evolution of drug-resistant malaria: The role of drug elimination half-life. *Philosoph Trans Biol Sci* 357, 505–519.
- Hotez PJ, and Brown AS. (2009). Neglected tropical disease vaccines. *Biologicals* 37, 160–164.
- Ivens AC, Peacock CS, and Worthey EA, et al. (2005). The genome of the kinetoplastid parasite, *Leishmania major*. *Science* 309, 436–442.
- Katoh K, Kuma K, Toh H, and Miyata T. (2005). MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33, 511–518.
- Koonin EV. (2005). Orthologs, paralogs, and evolutionary genomics. *Ann Rev Genetics* 39, 310–338.
- Li L, Stoeckert CJ, and Roos D. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* 13, 2178–2189.
- Li W, and Godzik A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659.
- Lindoso JAL, and Lindoso AABP. (2009). Neglected tropical diseases in Brazil. *Revista Instituto Med Tropical São Paulo* 51, 247–253.
- Loftus B, Anderson I, Davies R, et al. (2005). The genome of the protist parasite *Entamoeba histolytica*. *Nature* 433, 865–868.
- Lorenzi H, Puiu D, Brinkak L, and Amedeo P. (2010). New assembly, reannotation and analysis of the *Entamoeba histolytica* genome reveal new genomic features and protein content information. *PLoS Neglected Tropical Dis* 4, e716.
- Mulder NJ, Apweiler R, Attwood TK, et al. (2005). InterPro, progress and status in 2005. *Nucleic Acids Res* 33, D201–D205.
- Nobrega MA, and Pennacchio LA. (2004). Comparative genomic analysis as a tool for biological discovery. *J Physiol* 554, 31–39.
- Ocaña KCS, and Dávila AMR. (2011). Phylogenomics-based reconstruction of protozoan species tree. *Evolutionary Bioinformatic Online* 7, 107–121.
- Ofran Y, Punta M, Schneider R, and Rost B. (2005). Beyond annotation transfer by homology: Novel protein-function prediction methods to assist drug discovery. *Drug Discovery Today* 10, 1475–1482.
- Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, and Kanehisa M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 27, 29–34.
- Peacock CS, Seeger K, Harris D, et al. (2007). Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nature Genet* 39, 839–847.
- Reeves GA, Talavera D, and Thornton JM. (2009). Genome and proteome annotation: Organization, interpretation and integration. *J Royal Soc* 6, 129–147.
- Salavati R, and Najafabadi HS. (2010). Sequence-based functional annotation: What if most of the genes are unique to a genome? *Trends Parasitol* 26, 225–229.
- Shateri Najafabadi H, and Salavati R. (2010). Functional genome annotation by combined analysis across microarray studies of *Trypanosoma brucei*. *PLOS Neglected Trop Dis* 4, e810.
- Sonnhammer ELL, and Koonin EV. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* 18, 619–620.
- Tatusov R, Fedorova N, Jackson J, and Jacobs A. (2003). The COG database: An updated version includes eukaryotes. *BMC Bioinform* 4, 41.
- Trouiller P, Olliaro P, Torreele E, et al. (2002). Drug development for neglected diseases: A deficient market and a public-health policy failure. *Lancet* 359, 2188–2194.
- Weedall GD, and Hall N. (2011). Evolutionary genomics of *Entamoeba*. *Res Microbiol* 162, 637–645.
- Wilson CA, Kreychman J, and Gerstein M. (2000). Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* 297, 233–249.
- Ximénes C, Cerritos R, Rojas L, and Dolabella S. (2010). Human amebiasis: Breaking the paradigm. *Intl J Environ Res Public Health* 7, 1105–1120.
- Xu P, Widmer G, Wang Y, et al. (2004). The genome of *Cryptosporidium hominis*. *Nature* 431, 1107–1112.
- Yutin N, Makarova KS, Mekhedov SL, Wolf YI, and Koonin EV. (2008). The deep archaeal roots of Eukaryotes. *Mol Biol Evolution* 25, 1619–1630.

Address correspondence to:

Dr. Alberto M. R. Dávila
Computational and Systems Biology Laboratory
Computational and Systems Biology Pole
Oswaldo Cruz Institute
Fiocruz, Avenida Brasil
4365 Rio de Janeiro
Brazil

E-mail: davila@fiocruz.br