Molecular Evolution of the Vesicle Coat Component βCOP in Toxoplasma gondii

Sherri S. Smith[1, 2], *Stacy L. Pfluger* [1, 2, 3], *Elizabeth Hjort*[1],
Andrew G. McArthur[4], *Kristin M. Hager* [1, 2*]

1. Department of Biological Sciences, 2. Center for Global Health and Infectious Disease, University of Notre Dame, Notre Dame, IN 46556-0369 USA. 50011 USA. 3. Current Address: Department of Biology, Baylor University, Waco, Texas, 76798-7388. 4. Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA

Running Title: Functional Divergence of βCOP.

Abbreviations: coatomer protein complex I (COPI), beta-COP (βCOP), ER (endoplasmic reticulum), ORF (open reading frame),

Keywords: coatomer; beta-COP (β-COP); phylogeny; Apicomplexa; vesicle trafficking; DIVERGE; functional divergence.

*Author for correspondence (e-mail: khager1@nd.edu)

*Corresponding author:
Dr. K.M. Hager
Department of Biological Sciences
Center for Global Health and Infectious Disease
216 Galvin Life Sciences Building
University of Notre Dame
Notre Dame, IN. 46556-0369 USA

ABSTRACT

Coatomer coated (COPI) vesicles play a pivotal role for multiple membrane trafficking steps throughout the eukaryotic cell. Our focus is on βCOP, one of the most well known components of the COPI multi-protein complex. Amino acid differences in βCOP may dictate functional divergence across species during the course of evolution, especially with regards to the evolutionary pressures on obligate intracellular parasites. A bioinformatic analysis of βCOP amino acid sequences was conducted for 49 eukaryotic species. Cloning and sequence analysis of the *Toxoplasma gondii* βCOP homologue revealed several amino acid insertions unique to *T. gondii* and one C-terminal insertion that is unique to apicomplexan parasites. These findings led us to investigate the possibility that βCOP experienced functional divergence during the course of its evolution. Bayesian phylogenetic analysis revealed a tree consistent with pan eukaryote distribution and long-branch lengths were observed among the apicomplexans. Further analysis revealed that kinetoplast βCOP underwent the most amount of change, leading to perhaps an overall change of function. In comparison, *T. gondii* exhibited subtle yet specific amino acid changes. The amino acid substitutions did not occur in the same places as other lineages, suggesting that *Tgβ**COP*** has a role specific to the apicomplexans. Our work identifies forty-eight residues that are likely to be functionally important when comparing apicomplexan, kinetoplastid, and fungal βCOP. Keywords: coatomer; beta-COP (β-COP); phylogeny; Apicomplexa; vesicle trafficking; DIVERGE; functional divergence.

1. INTRODUCTION

Toxoplasma gondii is an obligate intracellular parasite and unicellular eukaryote. *T. gondii* uses conventional membrane trafficking organelles and machinery, such as the Endoplasmic Reticulum (ER) and coated vesicles, to accommodate a sophisticated secretory pathway that includes three additional secretory organelles utilized during invasion (Hager et al., 1999b; Joiner and Roos, 2002; Pfluger et al., 2005). The anterior cytoplasm of the parasite contains two of the specialized secretory organelles, called micronemes and rhoptries. This complex of specialized substructures gives the phylum Apicomplexa its name (Levine, 1977). Several apicomplexans also possess another distinctive organelle apical to the nucleus, the "apicoplast" (Abrahamsen et al., 2004a; Köhler et al., 1997; Lang-Unnasch et al., 1998; McFadden et al., 1996; Zhu et al., 2000). This organelle is thought to have evolved through secondary endosymbiosis and is plastid in nature (Kohler et al., 1997; Waller and McFadden, 2005). Unlike other apical organelles, it harbors its own genome and is enclosed within four membranes (Kohler et al., 1997; McFadden and Roos, 1999).

Other than the nucleus, the presence of an endomembrane system within the cytoplasm of an organism is one of the hallmarks of eukaryotes. In most eukaryotic cells, proteins move from one compartment to another in transport vesicles. These transport vesicles form from specially coated regions of membranes and bud off as coated vesicles, with a distinctive cage of proteins covering the surface facing the cytosol. There are two well-characterized types of transport vesicles: clathrin-coated and coatomer-coated (COP) vesicles. Clathrin mediates selective transport of transmembrane receptors from the trans-Golgi network or from the plasma membrane. COP vesicles mediate both selective and non-selective transport between the ER and Golgi and/or within the Golgi cisternae (Barlowe, 2000; Cox, 2002; Majoul et al., 2001; Nickel et al., 2002; Orci et al., 2000; Pelham and Rothman, 2000; Tang et al., 2001)[13-19]. There are two types of COP vesicles, COPI and COPII. Direction and control of

early transport within the secretory pathway is mediated by the components of COPI. COPI contains stoichiometric amounts of seven subunits including: α(160-kDa,) β (107-kDa), β' (102-kDa), δ (57-kDa), ε (36-kDa), γ (97-kDa), ζ (20-kDa) and a small GTP-binding protein, ADP-ribosylation factor (ARF). (Lowe and Kreis, 1998)[20]. COPI mediates retrograde trafficking that involves retrieval of missorted proteins from the Golgi back to the endoplasmic reticulum (ER). Less well understood, and a current subject for debate, are COPI's other roles in the cell. COPI has been implicated in anterograde stack to stack transport (Orci et al., 2000), concentration of proteins at specialized ER exit sites (ERES) where cargo leaves the ER (Garcia-Mata et al., 2003), endocytosis (Aniento et al., 1996) and phagosome maturation (Botelho et al., 2000). In contrast to COPI, COPII's role in transport is well understood. It is required for the concentration and export of secretory cargo from the ER (Aridor et al., 1999; Barlowe, 1998; Barlowe, 2000; Gorelick and Shugrue, 2001; Gurkan et al., 2006). COPII contains four subunits (different in sequence from the COPI subunits), that includes a small GTP-binding protein, SAR1. With the exception of the small GTPases, all COPI and COPII subunits share an 'Adaptin_ N' domain, at the N-terminus of all of the coatomer subunit proteins. Adaptin domains are present in the alpha, beta and gamma subunits of the AP-1, AP-2 and AP-3 adaptor protein complexes. The adaptor protein (AP) complexes are involved in the formation of clathrin coated pits and vesicles. The N-terminal region of adaptor proteins is conserved in sequence compared to the C-terminal region that varies within the family. The C-terminal region has been proposed to interact with other components of the coated vesicle and aid in budding of the vesicle (Antonny, 2006; Gurkan et al., 2006).

The *Toxoplasma* genome possesses all seven of the predicted COPI subunit homologues and the predicted sizes are: α (142.6 kDa), β (121 kDa), β' (150...154 kDa), δ (62 kDa), ε (34 kDa) , γ (112 kDa), ζ (21 kDa). One of the best known constituents of the COPI complex is βCOP. We conducted a bioinformatic analysis on βCOP because it is essential for membrane trafficking steps throughout the

eukaryotic cell and its role in trafficking is an active area of research. We have recently cloned and characterized the *Toxoplasma* homologue of βCOP (TgβCOP) and found that it has several conserved and unique properties. It localizes to the parasites' Golgi region (Hager et al., 1999; Pfluger et al., 2005). TgβCOP is slightly larger than other metazoan or fungal βCOP homologues. *Tg*βCOP possesses the canonical 'Adaptin_ N' domain at its N-terminus and also contains an amino acid insert in its C-terminal domain that is conserved across several apicomplexan species, but is not found in other sequenced metazoan or fungi βCOP. This study examines the phylogenetic relatedness of TgβCOP to other eukaryotes in order to identify regions that may be associated with divergent biological functions.

2. MATERIALS AND METHODS

2.1. Reagents and Cell Culture

All chemical reagents and primers were ordered as previously described (Pfluger et al., 2005). All cell culture reagents were obtained from Gibco (Invitrogen, Carlsbad, CA). The host cells, human foreskin fibroblast (HFF) cells (BJ cell line, ATCC), and parasites were grown as previously described (Roos et al., 1994).

2.2. Identification and cloning of TgβCOP

An expressed sequence tag (EST) encoding βCOP (accession #**AF163574**) (Hager et al., 1999a) was identified from the *Toxoplasma* EST Database (Ajioka et al., 1998; Kissinger et al., 2003; Li et al., 2003; www.ToxoDB.org, 2002). The putative ORF was initially identified using primers directed against the 5' sequences in the EST sequence #**623080** using 5' RACE PCR. In brief, the five prime end of *T. gondii* EST #**623080** was used to design primer *Tg*βCOP1 (Rev). TgβCOP gene specific primer (GSP) in the antisense orientation was used with the universal primer (sense) in a 5'-RACE

PCR strategy according to manufacturers protocol (Stratagene). The resulting PCR product was sub-cloned into pCR4-TOPO vector (Invitrogen) and sequenced using standard T3 and T7 primers. To confirm the sequence of PCR product, TgβCOP, primers were designed to the ends of the gene (*Tg*βCOP2 and *Tg*βCOP3) and to an interior region of the gene (*Tg*βCOP4 and *Tg*βCOP5) based on the previously determined sequence. The interior primers (*Tg*βCOP4 and *Tg*βCOP5) were also used in conjunction with the 5' and 3' end primers (*Tg*βCOP2 and *Tg*βCOP3) to produce PCR fragments of appropriate lengths to re-confirm cloning and sequencing result. The complete ORF was cloned from full-length cDNAs generated by using SMART cDNA library construction (Becton Dickinson) from the RH strain. All sequencing was performed by Davis Sequencing Facility (Davis, CA). Cloning and sequencing primers are shown in Table I.

2.4. Phylogenetic and functional analysis

Sequences of βCOPs from 49 species were obtained by searching NCBI's non-redundant proteins database and protozoan databases using BLASTP. The sequences obtained were aligned using default parameters in ClustalX (Thompson et al., 1997). Manual correction was used to exclude gaps and highly divergent or ambiguous regions of the alignment. Therefore, only the conserved core of the alignment (black portions, Figure 1A) was used to generate the phylogenetic tree shown in Figure 2. TgβCOP sequence corresponding to amino acids 990-1043 shaded region (See Fig. 1A, insert) was aligned with other species utilizing the program T-Coffee (www.ch.embnet.org/software/TCoffee.html) (Notredame et al., 2000). Abbreviations and accession numbers used in the alignment inset in Fig. 1A are as follows: *Hs35, Homo sapien* (NP_057535)*; Dm00, Drosophila melanogaster* (NP_523400)*; Ce36, Caenorhabditis elegans* (AAD12836)*; At77, Arabidopisis thalia* (NP_194877)*; Dd79, Dictystelium Discoidum (*AAF62179*)*; *Sc24, Saccharomyces cerevisiae* (CAA89724); *Eh34, Entamoeba histolytica* (EAL49134)*; Tb00, Trypanosoma brucei* (CAB95500)*; Gl00, Giardia lamblia* (EAA41300); *Cp14, Cyrptosporidium parvum*

(AAEE01000014*); Pf66, Plasmodium falciparum* (NP_702166); *Pr05, P. reichnow (*13d05.q1k, tBLASTN, genome, SANGER*); Pg03, P. gallinaceum* (l0373a03.q1k, tBLASTN, genome, SANGER); *Pk70, P. knowlesi (*14_1270w, BLASTP, protein, SANGER*); Pv50, P. vivax (*SaI-1|ctg_6877|Pv085050, BLASTP, protein, TIGR); *Py65, P. yoelli (*EAA20565*); Pb47, P. baudin* (XP_676947)*; Pc34, P. chaubadi* (CAH78234)*; Nc92, Neospora cannium* (CD667392)*; Tg21, Toxoplasma gondii* (DQ279721). Note, in cases where partial gene sequences are used, gene identifiers, the blast program and the database they were obtained from, are included within the parenthesis.

Phylogenetic relationships of the βCOP sequences were assessed using the computer program Mr. Bayes, a Bayesian statistical program (Huelsenbeck et al., 2001). Mr. Bayes works on the principle of Metropolis-Coupled Markov chain Monte Carlo (MC^3) estimation of posterior probabilities (Huelsenbeck et al., 2002; Lewis, 2001a; Lewis, 2001b; Shoemaker et al., 1999). We performed MC^3 estimation of posterior probabilities using noninformative prior probabilities, the JTT+I+Γ (Jones et al., 1992) substitution model with inclusion of unequal amino acid frequencies, and four incrementally heated Markov chains with different random starting trees. JTT was used as it is based on a broad sample of protein diversity and thus reflects empirical patterns of amino acid substitution. The Metropolis-coupled Markov chains were run to 10,000,000 generations with sampling every 100 generations. Posterior probabilities of topologies, clades, and parameters were estimated from the sampled topologies after removal of MC^3 burn-in. Four separate MC^3 analyses were performed to evaluate stationarity.

A functional divergence analysis of the amino acid alignment in the context of the hypothesized phylogenetic tree was performed using the computer program DIVERGE (Gu and Vander Velden, 2002). These sites may have been subject to divergent functional evolution. DIVERGE measured change in site-specific evolutionary rates using the coefficient of evolutionary functional divergence (θ), where $\theta = 0$ indicates no change and values approaching $\theta = 1$ reflect increasing functional

divergence. Specifically, θ measured site-specific divergence from a homogenous gamma model of among-site rate variation. Surface probabilities of the amino acid residues in TgβCOP sequence were calculated using the MacVector program (Rastogi, 2000) obtained from the website: www.macvector.com/index.html (MacVector, Inc.). The program was run using default parameters.

2.5. Database acknowledgements

Preliminary genomic, EST and/or cDNA sequence data for *T. gondii* was accessed via http://ToxoDB.org and/or http://www.tigr.org/tdb/t_gondii/. EST data for *Neospora caninum* was obtained from NCBI. Genomic data for Plasmodium species were provided by The Institute for Genomic Research (supported by the NIH grant #AI05093), and the Sanger Center (Wellcome Trust). EST sequences were generated by Washington University (NIH grant #1R01AI045806-01A1). Sequence data for *P. falciparum* chromosome 14 was obtained from The Sanger Institute website at http://www.sanger.ac.uk/Projects/P_falciparum/. Sequencing of *P. falciparum* chromosome 14 was accomplished as part of the Malaria Genome Project with support by The Wellcome Trust. See Table 2 for accession number and species used in Bayesian and DIVERGE analysis.

3. Results And Discussion

3.1. cDNA cloning TgβCOP and sequence analysis. The full length cDNA for the TgβCOP gene (accession # **DQ279721**) is 3.3 kb (Figure 1B) in size and contains overlapping sequence with the EST fragment previously identified (Hager et al., 1999b). The coding region is predicted to encode a protein with a predicted mass of ~121-kDa. TgβCOP possesses a series of novel sequences distributed

throughout the protein (Fig. 1A, white boxes) that have been verified by reverse transcription polymerase chain reaction (RT-PCR) (Fig 1B). One of the novel sequences, the region between amino acids ~746-790 (using anti-GEDSL" antibody) exists in the mature protein as shown by immunoblotting (*Smith et al 2006 submitted*). The novel sequences have no identifiable function and/or motif. A novel sequence in the C-termini is shared among related Apicomplexa (Fig. 1A, striped box). While the sequence is not identical between all apicomplexans, a shared amino acid motif suggests common ancestry (Fig. 1A, alignment). This novel sequence led us to investigate the role evolution played in exerting selective pressure for divergent function of the βCOP subunit among intracellular pathogens.

3.2. Phylogenetic analysis

A phylogenetic analysis of βCOP amino acid sequences provides a way to factor out historical and neutral processes and determine regions undergoing positive selection for functional divergence among eukaryotes. TgβCOP was compared with βCOP sequences representing 49 species of eukaryotes. The sequence alignment was modified to exclude regions of uncertain homology, a pre-requisite for accurate phylogenetic reconstruction. The final alignment included 825 characters. Using the final alignment, a consensus tree was generated by running MrBayes for 10 million MC^3 generations and excluding the burn-in trees (Figure 2). All four replicate MC^3 analyses produced the same results, indicating stationarity was obtained. Individual posterior probabilities varied no more than ±0.02. In each replicate analysis, the burn-in was 6000 generations, resulting in a sample of 99,940 trees for estimation of posterior probabilities.

The components of protist coated vesicular apparatus are of particular interest as these organisms often lack easily recognizable secretory organelles such as the Golgi. For example, *Giardia* (Gl) possesses an endomembrane system, yet appears to lack a conventional Golgi with readily recognizable and

parallel cisternae apparatus in its vegetative trophozoite form (Becker and Melkonian, 1996; Hehl and Marti, 2004; Lujan et al., 1995) despite an ability to sort proteins to constitutive and regulated secretory pathways (Meng et al., 1993; Reiner et al., 1990). We observed long branch lengths within the Apicomplexa clade (Figure 2), in excess of those observed in ribosomal RNA (rRNA) phylogenies (Sogin and Silberman, 1998). Long branches were observed in the other protist lineages and these may reflect a combination of time and elevated rates of molecular evolution. As such, the long branches observed may, in part, indicate functional divergence among the different protistan βCOP proteins, although sampling of distantly-related paralogs cannot be ruled out. In phylogenetic reconstruction, one of the most important tools to accurately predict branch lengths is use of a gamma distribution of among-site variation in evolutionary rates of change, as we have done in our MC^3 analysis. Our DIVERGE prediction of functionally divergent sites looks for exceptions to this distribution. While we recognize that comparison of βCOP branch-lengths to rRNAs may not be ideal as the apicomplexan SSU-rRNA family is small (Abrahamsen et al., 2004b; Gardner et al., 2005; Gardner et al., 2002; Pain et al., 2005; Xu et al., 2004) and may be under selection (personal communication, Jessica Kissinger), especially since stage-specific expression is observed in *Plasmodium* (Li et al., 1997; McCutchan et al., 1995), our use of DIVERGE compares rates of βCOP evolution between the Apicomplexa and other groups of organisms. Completely accurate reconstruction of Apicomplexa relationships is not essential. Accurate reconstruction of apicomplexan relationships may be better calculated using multiple protein encoding genes instead of our analysis of βCOP alone (Baldauf, 2003).

3.3. DIVERGE analysis

To examine the possibility of functional divergence of apicomplexan βCOP proteins, we performed a functional analysis of the amino acid alignment in the context of the βCOP phylogenetic tree using the computer program DIVERGE (Gu and Vander Velden, 2002). DIVERGE detects site-specific altered

functional constraints by comparing site-specific evolutionary rates among sub-clades within the phylogenetic tree (reviewed by Gaucher *et al.*, 2002)(Gaucher et al., 2002). DIVERGE does not explicitly consider the nature of amino acid substitutions occurring at specific locations in the alignment (e.g. non-polar vs. polar vs. charged polar), but instead highlights positions that have higher rates of amino acid substitution than predicted from neutrality. As such, DIVERGE does not highlight extreme conservation, such as structurally important cysteine residues, but only those positions within the alignment undergoing statistically detectable positive selection. DIVERGE performs the analysis by comparing the evolutionary rates between two clades within the phylogenetic tree. The tree is used to factor out neutral processes by using a mathematical model of the rates of evolution within a protein. Sites having significantly higher evolutionary rates among sub-clades than predicted from neutrality are identified as regions potentially subject to divergent functional evolution. However, it is important to note that the unique population structure of parasites such as the apicomplexa and kinetoplastid may additionally contribution to rates of evolution beyond that expected from neutrality. As such, our results should be considered hypothesis testable by functional studies in the laboratory.

We used DIVERGE to test the null hypothesis of no changes in site-specific evolutionary rates among βCOP sub-clades and to predict sites in the alignment having altered functional constraints. Specifically, we tested this null hypothesis among the apicomplexans, kinetoplastids, vertebrates, and fungi (Table 3). The kinetoplastids exhibited a high degree of functional divergence from the vertebrates ($\theta = 0.66$ overall, 577 residues with $\theta > 0.50$) and fungi ($\theta = 0.46$ overall, 181 residues with $\theta > 0.50$), indicating broad positive selection on the amino acid sequence between kinetoplastids and vertebrates / fungi. As DIVERGE is strictly comparative, it is not explicitly clear if this selective pressure has been at work for the kinetoplastid proteins only, for the vertebrate and fungal proteins only, or a combination of each. However, examination of Table 3 finds no evidence of positive selection between the vertebrate and fungal proteins, suggesting that the positive selection is occurring

predominantly in the kinetoplastids. Kinetoplastid βCOP may be performing a very novel function compared to well-studied model organisms. Previous studies showed that trypanosome βCOP antibodies do not label the Golgi (Maier et al., 2001) and this may possibly be related to their high level of divergence.

The apicomplexans exhibited a subtle and specific functional divergence from fungi ($\theta = 0.30$, 28 residues with $\theta > 0.50$) and kinetoplastids ($\theta = 0.35$, 23 residues with $\theta > 0.50$), but not vertebrates. Apicomplexans, unlike kinetoplastids, contain a small number of residues predicted to have undergone positive selection (< 30 for each comparison). Although the divergent residues were found throughout the protein, the majority were clustered in the conserved N-terminal and C-terminal domains, with many functionally divergent residues found in close proximity (Figure 3, Table 4). Only three divergent positions were shared between apicomplexans, kinetoplastids, and fungi. Despite being in close proximity, some positions undergoing positive selection appear to be divergent in the apicomplexans only, the kinetoplastids only, or the fungi only (Table 4). Sites where non-conservative amino acid substitutions occur are of interest. For example, at position 133 most fungi have a glutamine (an uncharged polar amino acid and amide), while apicomplexans have a proline (a non-polar amino acid and a bulky aliphatic hydrocarbon). This substitution of proline can result in disruption of an alpha helix, a beta-sheet or change the direction of a beta-strand within a protein (Table 4).

The topological position of the residues identified in Figure 3 (Table 4) is important to consider for functional divergence for two reasons. First, residues that are predicted to be on the surface generally interact favorably with the hydrophilic environment of the cytosol through polar amino acid groups. Second, surface residues would presumably have the most impact on protein-protein interactions within the cytosol of the organism. Given that βCOP is part of a multi-protein complex, the

implications are that these interactions would change dependent upon the type of amino acid substitutions that are observed. When those groups are changed to non-polar residues (for example at position 603 or 941), the interactions with the hydrophilic environment of the cytosol is less favored. The result is that in order to maximize hydrogen bonding, the protein may change how it folds. Thus, non-conservative amino acid changes can affect the overall shape of the protein and presumably the protein-protein interactions that occur between the protein and its subunits within the coatomer complex. We hypothesize that each of these lineages have been undergoing subtle divergent evolution to fine tune their roles in their respective cells and that the DIVERGE analysis has highlighted regions of the protein important for alteration of protein structure and function. That each group of organisms did not undergo amino acid substitutions at the exact same positions is strongly suggestive of differing selective pressures. As such, TgβCOP may play a role specific to the apicomplexans. The three divergent sites shared between apicomplexans, fungi, and kinetoplastids are not in the conserved domains of the coatomer beta subunit proteins, suggesting previously undiscovered functionally important residues in βCOP. These residues, as well as the clusters of divergent residues, are important for future experimental investigation of βCOP function.

Acknowledgements

KMH is an Ellison Medical Foundation New Scholar in Infectious Disease; SLP was supported by an NIH training grant (NIH/NIAID/TMP T 32 7030); AGM was supported by the Marine Biological Laboratory's Program in Global Infectious Diseases, also funded by the Ellison Medical Foundation. Computational resources were provided by the Josephine Bay Paul Center for Comparative Molecular Biology and Evolution (Marine Biological Laboratory) through funds provided by the W.M. Keck Foundation and the G. Unger Vetlesen Foundation. We thank Jessica Kissinger and Jim Ajioka for helpful conversations regarding this work.

Figure Legends

Fig. 1. TgβCOP cloning and sequence analysis. (A) Schematic of TgβCOP. White boxes represent sequences with low conservation; the black boxes represent areas of high sequence conservation. Domains are marked as follows: N-termini=NH_2, adaptin-like domain found at amino(N)-terminus=Adaptin_N, the middle or hinge region of the protein=hinge, carboxyl(C)-terminus=COOH. This schematic was generated after alignment analysis against of 49 other eukaryotic species using default parameters in ClustalW. Sequences of βCOPs from multiple species were obtained by searching NCBI and protozoan databases (database acknowledgments and accession numbers are in Materials and Methods). The shaded sequence represents the novel C-terminal sequence shared across several apicomplexan species. The alignment for this region is shown the below schematic (generated using T-Coffee). Asterisks denote putative phosphorylation sites. (B) Reverse transcription (RT)-polymerase chain RT-PCR reaction (PCR) of TgβCOP. Lane one (1) are the DNA molecular weight markers. Lane two (2) is the 3.3 kb product (arrow) is shown, which results from RT-PCR using both 5' and 3' primers specific for generating the full-length cDNA encoding TgβCOP. (primers: TgβCOP alone (5`) and TgβCOP cDNA synthesizer, see materials and methods, and Table 1).

Fig. 2. Phylogenetic tree of βCOP species.

The conserved core of the βCOP alignments (indicated in the model by the black boxes, Figure 1A) was used to generate the phylogenetic tree. Phylogenetic relationships of the βCOP sequences were assessed using the computer program MrBayes, a Bayesian statistical program (Huelsenbeck et al., 2001). Branch lengths indicate evolutionary distance while node labels indicate posterior probability. The vertebrate and fungal branches were magnified (see inset boxes) to show relationships more clearly (branch lengths are not meaningful). Species abbreviations, their accession numbers, and percent identity to human βCOP (**NP_057535**) are shown in Table 2.

Fig. 3. Predicted functionally divergent residues in *Toxoplasma gondii*.

(A) Residues predicted to be functionally divergent between the Apicomplexa and the kinetoplastida or fungi are presented in bold. Clusters of divergent residues are shown as groups and are presented in their relative position within the βCOP homologue with respect the N and C-termini. Hinge region denotes the middle region of the protein. Black bolded residues are polar, black, bolded and underlined residues are charged polar amino acids. Red residues are hydrophobic and purple are sulfur-containing amino acids. (B) (Shown in supplement) The complete sequence of TgβCOP is shown, aligned with regions of fungi and kinetoplastida sequences used in the phylogenetic and divergence analyses. Regions of poor alignment are not shown. The conserved Adaptin N-terminal region and the conserved C-terminal domain of coatomer beta subunit proteins are underlined in the TgβCOP sequence.

Table 1. Oligonucleotide and gene-specific primers (GSP) used for reverse transcriptase-polymerase chain reaction (RT-PCR), rapid amplification of cDNA ends (RACE), and DNA sequencing.

Table 2. GenBank accession numbers for βCOP amino acid sequences used to generate multiple sequence alignments and phylogenetic trees, with identity to human βCOP.

Table 3. Test of the null hypothesis of no change in site-specific evolutionary rates among four clades of βCOP (apicomplexans, kinetoplastids, vertebrates, and fungi), as implemented by the computer program DIVERGE. Measures of the coefficient of evolutionary functional divergence (θ) and their standard error are presented above the diagonal. Values of θ above 0.50 up to 1.00 reflect increasing functional divergence. Likelihood ratio test observed values for the test of the null hypothesis of $\theta = 0$

are presented below the diagonal. Scores with significant rejection ($P < 0.05$) of the null hypothesis are marked with an asterisk.

Table 4. Table of divergent residues across kinetoplasts, fungi and Apicomplexa and their respective probabilities of being exposed the surface of the folded protein. Examples of invariant residues as they compare across species are shown in red. Variable means that the residue could be polar (uncharged or charged) or non-polar. Variable charge means that the residue could possess a positive, negative, or not charge at all. Polar=hydrophilic. Non-polar=hydrophobic. +=glycine possesses no optical isomer, and thus is in a category by itself. Note: DIVERGE does not take into account the type of amino acid but rather the changes that occur at that particular position within the primary amino acid sequence.

References

Abrahamsen, M. S., Templeton, T. J., Enomoto, S., Abrahante, J. E., Zhu, G., Lancto, C. A., Deng, M., Liu, C., Widmer, G., Tzipori, S. et al. (2004a). Complete genome sequence of the apicomplexan, Cryptosporidium parvum. *Science* **304**, 441-5.

Abrahamsen, M. S., Templeton, T. J., Enomoto, S., Abrahante, J. E., Zhu, G., Lancto, C. A., Deng, M., Liu, C., Widmer, G., Tzipori, S. et al. (2004b). Complete Genome Sequence of the Apicomplexan, Cryptosporidium parvum. *Science* **304**, 441-445.

Ajioka, J. W., Boothroyd, J. C., Brunk, B. P., Hehl, A., Hillier, L., Manger, I. D., Marra, M., Overton, G. C., Roos, D. S., Wan, K. et al. (1998). Gene Discovery by EST Sequencing in *Toxoplasma gondii* Reveals Sequences Restricted to the Apicomplexa. *Genome Research* **8**, 18-28.

Aniento, F., Gu, F., Parton, R. G. and Gruenberg, J. (1996). An endosomal beta COP is involved in the pH-dependent formation of transport vesicles destined for late endosomes. *J Cell Biol* **133**, 29-41.

Aridor, M., Bannykh, S. I., Rowe, T. and Balch, W. E. (1999). Cargo can modulate COPII vesicle formation from the endoplasmic reticulum. *J Biol Chem* **274**, 4389-99.

Baldauf, S. L. (2003). The deep roots of eukaryotes. *Science* **300**, 1703-6.

Barlowe, C. (1998). COPII and selective export from the endoplasmic reticulum. *Biochim Biophys Acta* **1404**, 67-76.

Barlowe, C. (2000). Traffic COPs of the early secretory pathway. *Traffic* **1**, 371-7.

Becker, B. and Melkonian, M. (1996). The secretory pathway of protists: spatial and functional organization and evolution. *Microbiol Rev* **60**, 697-721.

Blom, N., Gammeltoft, S. and Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* **294**, 1351-62.

Botelho, R. J., Hackam, D. J., Schreiber, A. D. and Grinstein, S. (2000). Role of COPI in phagosome maturation. *J Biol Chem* **275**, 15717-27.

Cox, E. H. (2002). Picture story. Shedding new light on coatomer dynamics. *Nat Struct Biol* **9**, 413.

Garcia-Mata, R., Szul, T., Alvarez, C. and Sztul, E. (2003). ADP-Ribosylation Factor/COPI-dependent Events at the Endoplasmic Reticulum-Golgi Interface Are Regulated by the Guanine Nucleotide Exchange Factor GBF1. *Mol Biol Cell* **14**, 2250-61.

Gardner, M. J., Bishop, R., Shah, T., de Villiers, E. P., Carlton, J. M., Hall, N., Ren, Q., Paulsen, I. T., Pain, A., Berriman, M. et al. (2005). Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes. *Science* **309**, 134-7.

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

Gaucher, E. A., Gu, X., Miyamoto, M. M. and Benner, S. A. (2002). Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem Sci* **27**, 315-21.

Gorelick, F. S. and Shugrue, C. (2001). Exiting the endoplasmic reticulum. *Mol Cell Endocrinol* **177**, 13-8.

Gu, X. and Vander Velden, K. (2002). DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics* **18**, 500-1.

Hager, K. M., Striepen, B., Tilney, L. G. and Roos, D. S. (1999a). The nuclear envelope serves as an intermediary between the ER and Golgi complex in the intracellular parasite Toxoplasma gondii. *J Cell Sci* **112**, 2631-8.

Hager, K. M., Striepen, B., Tilney, L. G. and Roos, D. S. (1999b). The nuclear envelope serves as an intermediary between the ER and Golgi complex in the intracellular parasite Toxoplasma gondii. *J Cell Sci* **112 (Pt 16)**, 2631-8.

Hehl, A. B. and Marti, M. (2004). Secretory protein trafficking in Giardia intestinalis. *Mol Microbiol* **53**, 19-28.

Huelsenbeck, J. P., Larget, B., Miller, R. E. and Ronquist, F. (2002). Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst Biol* **51**, 673-88.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology
MRBAYES: Bayesian inference of phylogenetic trees. *Science* **294**, 2310-4.

Joiner, K. A. and Roos, D. S. (2002). Secretory traffic in the eukaryotic parasite Toxoplasma gondii: less is more. *J Cell Biol* **157**, 557-63.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275-82.

Kissinger, J. C., Gajria, B., Li, L., Paulsen, I. T. and Roos, D. S. (2003). ToxoDB: Accessing the *Toxoplasma gondii* genome. *Nucl. Acids Res.* **31**, 234-236.

Kohler, S., Delwiche, C. F., Denny, P. W., Tilney, L. G., Webster, P., Wilson, R. J., Palmer, J. D. and Roos, D. S. (1997). A plastid of probable green algal origin in Apicomplexan parasites. *Science* **275**, 1485-9.

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

Lang-Unnasch, N., Reith, M. E., Munholland, J. and Barta, J. R. (1998). Plastids are widespread and ancient in parasites of the phylum Apicomplexa. *Int J Parasitol* **28**, 1743-54.

Levine, N. D. (1977). Taxonomy of Toxoplasma. *J Protozool* **24**, 36-41.

Lewis, P. O. (2001a). A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst Biol* **50**, 913-25.

Lewis, P. O. (2001b). Phylogenetic systematics turns over a new leaf. *Trends in Ecology and Evolution* **16**, 30-37.

Li, A. L., Brunk, B. P., Kissinger, J. C., Pape, D., Tang, P., Cole, R., Martin, J., Wylie, T., Dante, M., Fogarty, S. J. et al. (2003). Gene discovery in the Apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. *Genome Research* **13**, 443-454.

Li, J., Gutell, R. R., Damberger, S. H., Wirtz, R. A., Kissinger, J. C., Rogers, M. J., Sattabongkot, J. and McCutchan, T. F. (1997). Regulation and trafficking of three distinct 18 S ribosomal RNAs during development of the malaria parasite. *J Mol Biol* **269**, 203-13.

Lowe, M. and Kreis, T. E. (1998). Regulation of membrane traffic in animal cells by COPI. *Biochim Biophys Acta* **1404**, 53-66.

Lujan, H. D., Marotta, A., Mowatt, M. R., Sciaky, N., Lippincott-Schwartz, J. and Nash, T. E. (1995). Developmental induction of Golgi structure and function in the primitive eukaryote Giardia lamblia. *J Biol Chem* **270**, 4612-8.

Maier, A. G., Webb, H., Ding, M., Bremser, M., Carrington, M. and Clayton, C. (2001). The coatomer of Trypanosoma brucei. *Mol Biochem Parasitol* **115**, 55-61.

Majoul, I., Straub, M., Hell, S. W., Duden, R. and Soling, H. D. (2001). KDEL-cargo regulates interactions between proteins involved in COPI vesicle traffic: measurements in living cells using FRET. *Dev Cell* **1**, 139-53.

McCutchan, T. F., Li, J., McConkey, G. A., Rogers, M. J. and Waters, A. P. (1995). The cytoplasmic ribosomal RNAs of Plasmodium spp. *Parasitol Today* **11**, 134-8.

McFadden, G. I., Reith, M. E., Munholland, J. and Lang-Unnasch, N. (1996). Plastid in human parasites. *Nature* **381**, 482.

McFadden, G. I. and Roos, D. S. (1999). Apicomplexan plastids as drug targets. *Trends Microbiol* **7**, 328-33.

Meng, T. C., Hetsko, M. L. and Gillin, F. D. (1993). Antigenic switching of TSA 417, a trophozoite variable surface protein, following completion of the life cycle of Giardia lamblia. *Infect Immun* **61**, 5394-7.

Nickel, W., Brugger, B. and Wieland, F. T. (2002). Vesicular transport: the core machinery of COPI recruitment and budding. *J Cell Sci* **115**, 3235-40.

Notredame, C., Higgins, D. G. and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-17.

Orci, L., Ravazzola, M., Volchuk, A., Engel, T., Gmachl, M., Amherdt, M., Perrelet, A., Sollner, T. H. and Rothman, J. E. (2000). Anterograde flow of cargo across the golgi stack potentially mediated via bidirectional "percolating" COPI vesicles. *Proc Natl Acad Sci U S A* **97**, 10400-5.

Pain, A., Renauld, H., Berriman, M., Murphy, L., Yeats, C. A., Weir, W., Kerhornou, A., Aslett, M., Bishop, R., Bouchier, C. et al. (2005). Genome of the host-cell transforming parasite Theileria annulata compared with T. parva. *Science* **309**, 131-3.

Pelham, H. R. and Rothman, J. E. (2000). The debate about transport in the Golgi--two sides of the same coin? *Cell* **102**, 713-9.

Pfluger, S. L., Goodson, H. V., Moran, J. M., Ruggiero, C. J., Ye, X., Emmons, K. M. and Hager, K. M. (2005). Receptor for Retrograde Transport in the Apicomplexan Parasite Toxoplasma gondii. *Eukaryot Cell* **4**, 432-42.

Rastogi, P. A. (2000). MacVector. Integrated sequence analysis for the Macintosh. *Methods Mol Biol* **132**, 47-69.

Reiner, D. S., McCaffery, M. and Gillin, F. D. (1990). Sorting of cyst wall proteins to a regulated secretory pathway during differentiation of the primitive eukaryote, Giardia lamblia. *Eur J Cell Biol* **53**, 142-53.

Roos, D. S., Donald, R. G., Morrissette, N. S. and Moulton, A. L. (1994). Molecular tools for genetic dissection of the protozoan parasite Toxoplasma gondii. *Methods Cell Biol* **45**, 27-63.

Shoemaker, J. S., Painter, I. S. and Weir, B. S. (1999). Bayesian statistics in genetics: a guide for the uninitiated. *Trends Genet* **15**, 354-8.

Sogin, M. L. and Silberman, J. D. (1998). Evolution of the protists and protistan parasites from the perspective of molecular systematics. *Int J Parasitol* **28**, 11-20.

Tang, B. L., Ong, Y. S., Huang, B., Wei, S., Wong, E. T., Qi, R., Horstmann, H. and Hong, W. (2001). A membrane protein enriched in endoplasmic reticulum exit sites interacts with COPII. *J Biol Chem* **276**, 40008-17.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876-82.

Waller, R. F. and McFadden, G. I. (2005). The apicoplast: a review of the derived plastid of apicomplexan parasites. *Curr Issues Mol Biol* **7**, 57-79.

www.ToxoDB.org. (2002). ToxoDB: Toxoplasma Genome Sequencing Consortium.

Xu, P., Widmer, G., Wang, Y., Ozaki, L. S., Alves, J. M., Serrano, M. G., Puiu, D., Manque, P., Akiyoshi, D., Mackey, A. J. et al. (2004). The genome of Cryptosporidium hominis. *Nature* **431**, 1107-12.

Zhu, G., Marchewka, M. J. and Keithly, J. S. (2000). Cryptosporidium parvum appears to lack a plastid genome. *Microbiology* **146**, 315-21.