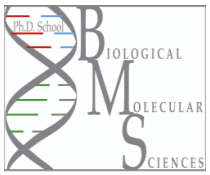


UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Scienze Biologiche e Molecolari

XXVII Ciclo

A systems biology approach to study the dynamics of membrane microdomains in malaria parasites.

Gabriella Sferra

PhD Thesis

Scientific tutors: Dr. Elisabetta PIZZI – Dr. Marta PONZI – Dr. Erica PASINI

Co-tutor: Dr. David HORNER

Academic year: 2013-2014

SSD: BIO/10 (Bioinformatica), VET/06 (Parassitologia)

Thesis performed at Department of Infectious, Parasitic and Immuno-mediated Diseases at the Istituto Superiore di Sanità, Rome.

Contents

PART I

PART II

Manuscript: The FAD-dependent glycerol-3-phosphate dehydrogenase of *Giardia duodenalis*: an unconventional enzyme that interacts with the g14-3-3 and it is a target of the antitumoral compound NDBHEX.
Authors: Marco Lalle, Serena Camerini, Serena Cecchetti, Renata Finelli, **Gabriella Sferra**, Joachim Müller, Giorgio Ricci, Edoardo Pozio
Journal: Frontiers in Microbiology, section Antimicrobials, Resistance and Chemotherapy
Accepted.

Part III

1 *Plasmodium berghei* and *Plasmodium knowlesi* phylogenetic profiling.
2 Phylogenetic profiling to predict host-parasite interactions in *P. falciparum* and *P. knowlesi*.
3 A multidomain protein from *Giardia duodenalis*, a rosetta stone case.

A systems biology approach to study the dynamics of membrane compartment in malaria parasites.

Gabriella SFERRA

PART I

Abstract

In recent years, several computational methods have been developed to predict protein-protein interactions at a genome-wide level. Among them, a Bayesian approach has been proposed to integrate "-omics" data from diverse sources, and reconstruct probabilistic global interactomes.

In order to apply this method to *Plasmodium falciparum*, the most virulent agent of the human malaria, we generated novel genomic data sets and gene expression profiles. In particular, we performed a re-assessment of the phylogenetic profile method proposing a new strategy to select reference genomes and adopting a novel measure of similarity. We also produced a new set of rosetta stone data on the basis of a large number of non-redundant genomes used as a reference set. Furthermore, diverse transcriptomic data have been organized to obtain gene expression profiles covering the entire intra-erythrocytic *Plasmodium* life-cycle. All data were then integrated to predict a global *P. falciparum* protein-protein interaction network. To gain insights on function and dynamics of specialized membrane compartments (lipid rafts), during *P. falciparum* development, we filtered our global interactome with stage-specific lipid raft proteomic data.

Functional and topological studies of the obtained stage-specific interactomes were undertaken. Our results revealed a conserved subnetwork, the lipid raft "functional core", involved in fundamental parasite processes and dynamic clusters populated of stage-specific proteins, responsible for remodeling of lipid raft organization.

1 – STATE OF THE ART

1.1 Prediction of protein-protein interaction networks.

Infectious diseases are the leading cause of human death, particularly in developing countries. Although in many cases drugs are available, the increasing emergence of pathogens resistant to drug treatments strongly requires to improve our understanding of the pathogen biology and in particular the key factors and processes that allow for disease establishment and progression. Biological processes occurring in a cell are driven by proteins that work in cooperation. For this reason, the understanding of a biological system requires a better knowledge of the interconnectivity between proteins in the different pathways and processes, which includes both physical and functional interactions.

In this perspective, reconstruction of protein-protein interaction (PPI) networks represents one of the challenges for both experimental and computational biologists. In the last years, several and diverse methods have been developed to infer PPIs. and different categories of predictors can be distinguished, as recently reviewed by Zahiri and colleagues (Zahiri et al., 2013).

Phylogenetic profiling is one of the most common methods to infer PPIs from genomic data, the basic idea being that functionally related proteins are encoded by coevolving genes and hence they may be captured by construction and comparison of their phylogenetic profiles. In the first implementation of the method, co-evolving proteins in a genome were identified by clustering binary

phylogenetic profiles reporting the presence (1) or the absence (0) of protein homologs in genomes of a reference set (Pellegrini, et al, 1999). To do this a BLAST data base search was carried out for each protein of a given organism *versus* the entire reference set. Hits with a value higher than an established threshold were regarded as possible homologs and registered as 1 in the profile. An implementation of the method was proposed by Date and collaborators substituting the binary profiles with real values derived from the E-values of the BLAST matches and adopting the mutual information metric to measure the profile similarity (Date, et al, 2003). Among the measures of profile similarity successively proposed, the correlation-based measures have been shown to have the best predictive performances even if they capture only linear relationships (Date, et al., 2003; Enault, et al., 2003; Glazko, et al, 2004).

Another computational method based on genomic data to predict PPIs is the rosetta stone fusion protein method. The method relies on the observation that certain interacting proteins can be predicted, due to the presence in other organisms of orthologs fused in a single gene, encoding a multi-domain product named “rosetta stone protein” (Enright, et al, 1999). Protein interactions are then identified by comparing the protein sequences of the genome under examination with those of a reference set, to look for gene fusion events (Enright, et al, 1999). A score is assigned to each predicted interaction, based on the probability that the corresponding fusion event occurred by chance (Marcotte & Marcotte, 2002).

The usage of gene expression data to infer PPIs is based on the observation that proteins participating to the same complex or involved in the same biological

pathway are often encoded by co-expressed genes (Jansen, et al, 2002). In the recent past, several technological platforms have been developed to measure the global gene expression in different cellular types. This led to the generation of a huge amount of expression data, available to the scientific community. Two methods are usually used to quantify simultaneously all transcripts of a biological sample: the microarray-based transcriptomics and, more recently, the direct RNA sequencing, also known as RNA-seq (Gibson, 2003; Nagalakshmi, et al, 2010). Co-expression of group of genes are established by means of classical statistical tools applied to time-course experiments.

However, PPIs predicted by different methods are not directly comparable because of the difference in their genome coverage and accuracy (Ito, et al, 2001; Deane, et al, 2002; Von Mering, et al, 2002). For these reasons, the integration of multiple datasets from different experimental technologies and platforms and from different computational methods becomes crucial to produce a comprehensive and accurate set of relationships between interacting proteins (Jansen, et al, 2003; Lee, et al, 2004).

Exploiting the Bayesian approach proposed by Jansen and colleagues (Jansen, et al, 2003) genomic and transcriptomic data sets, available for *Saccharomyces cerevisiae,* were used to derive a probabilistic PPI network. Each data set (phylogenetic profiles, gene-expression profiling, rosetta stone data) was compared with gold standard data sets, representing the benchmark, to derive a probability score. Under the assumption of independence of data sets, a global score was then obtained multiplying the single ones. It has been demonstrated that the performance of the integrated predictions strongly overcomes the one obtained from single approaches.

In this work, we applied the Bayesian method to derive a novel PPI network for the malaria parasite *Plasmodium falciparum* using phylogenetic profiles, rosetta stone and gene expression data. Taking advantage of the increasing number of sequenced genomes, a novel set of phylogenetic profiles and rosetta stone data were derived. In particular we proposed a new measure to score profile similarity based on a recently proposed metric of correlation between variables (Szekely, et al, 2007), the distance correlation.

1.2 Malaria, a global public health threat.

Malaria is one of the major global public health threat together with AIDS and tuberculosis. About 3.2 billion people worldwide are estimated to live in countries and territories, mainly corresponding to the tropics and subtropics regions, at risk of malaria transmission and disease. Morbidity and mortality deeply impact on these populations causing hundred millions of clinical cases and half million deaths yearly. The highest burden is focused in Sub-Saharian Africa: around 90% of all deaths occurs in this region and the greatest part of them affects children under the age of five (WHO, World Malaria Report 2014). Malaria transmission is strongly influenced by climatic features as temperature, humidity and rainfall (Gubler, et al, 2001), and in the last decades, the climate changes are modifying malaria spreading, making this parasitosis stable in regions at the margin of its previous distribution (Chaves, et al, 2010). Furthermore, socio-economic trends (i.e., migration and tourism) strengthen imported malaria cases increasing the risk of parasite transmission in temperate areas where the vectorial capacity is high, as in the Mediterranean area

(Odolini, et al, 2012). Moreover, the absence of an efficient vaccine and the growing number of resistance to drugs make necessary to improve our knowledge of the biology of the malaria parasites.

Malaria is caused by *Plasmodium*, a Protozoan belonging to the phylum of the Apicomplexa. *Plasmodium* life cycle takes place into two different hosts: a female mosquito of the genus *Anopheles* and a vertebrate host. Until recently, four species of the malaria parasite (*Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale* and *Plasmodium malariae*) were considered able to infect humans and between them *P. falciparum* is the most virulent agent of human malaria.

Parasite life cycle (figure 1) starts when an infected female mosquito of the genus *Anopheles* bites a vertebrate host. The sporozoites reach the liver where the first asexual replication takes place within the hepatocytes. In 9-16 days, each sporozoite that enters the host hepatocytes gives rise to thousands merozoites, the erythrocyte invasive form. Inside the erythrocyte, merozoites rapidly become larger, ring-shaped young trophozoites.

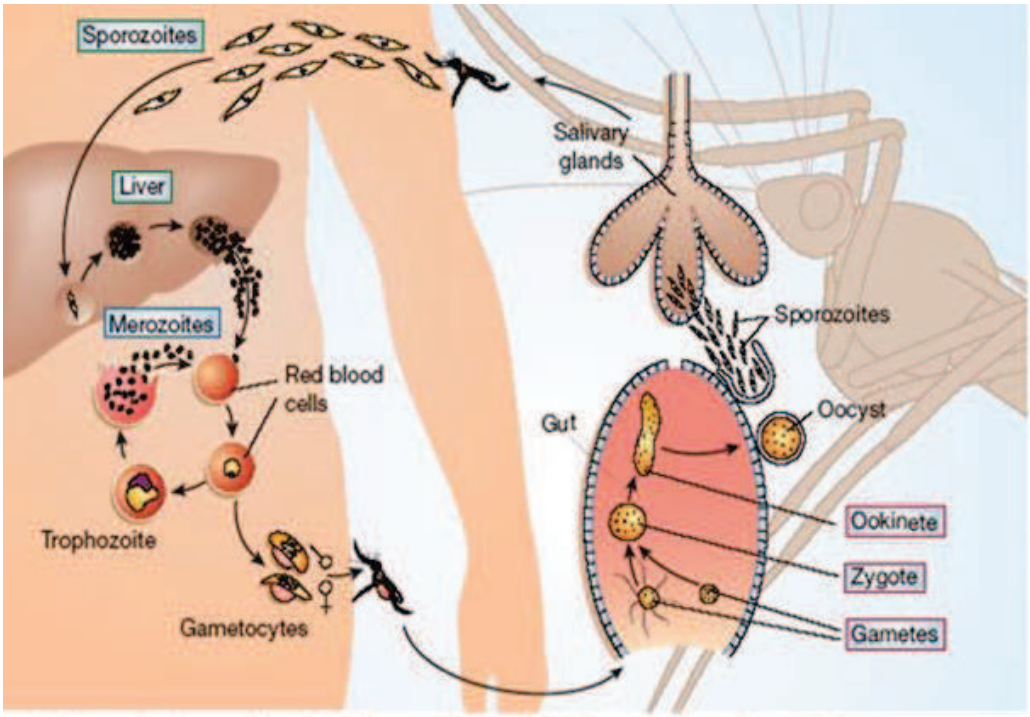


Figure 1. Plasmodium falciparum life cycle (from www. malwest.gr).

A rapid and repeated division of the trophozoite's nucleus forms a schizont, characterized by several nuclei immersed in the same cytoplasm. Following cytokinesis, 16-32 merozoites are formed per each erythrocyte. A pushing force coming from the inside-out turning of the erythrocyte membrane permits the egress of the merozoites from the infected red blood cell through a single osmotic-induced pore (Abkarian, et al, 2011). Red blood cell content is released together with the merozoites and is responsible of a large part of malaria symptoms. Released merozoites are ready to invade new erythrocytes. A small subpopulation of schizonts, committed to sexual differentiation (gametocytogenesis) differentiate into male and female gametocytes (Inselburg, 1983), the only forms infectious to mosquito (Smalley & Sinden, 1977).

1.3 Parasite research in the "–omic" era.

The complex biology of the parasite, the difficulties of retrieving the pure stages, its low transfection efficiency and its haploid state that limits classic genetic approaches or determines an excessive time cost (Balu, 2012) encourage the use of high-throughput approaches to characterize protein function at local and genome-wide level.

In 2002, the first *P. falciparum* genome was completely sequenced (Gardner, et al, 2002) opening a novel era for the understanding of the parasite biology. The availability of the complete genome sequence encouraged the application of high-throughput methods to the study of *Plasmodium* developmental stages. Along with the genome sequencing, the first proteomic analysis of different stages of parasite life cycle (Florens et al., 2002; Lasonder et al., 2002) was also provided. Several transcriptome analyses came soon and revealed the highly ordered gene expression cascade characteristic of *P. falciparum* (Le Roch, et al, 2003; Bozdech, et al, 2003). These data, along with the observation that only few genes encode for factors transcriptionally regulated, supported the hypothesis of a strong post-transcriptional control in *Plasmodium* gene expression (Llinas, 2006). More recently, massive parallel RNA-sequencing approaches (Bentley, et al, 2008) further improve the status of the genome annotation (Su, et al, 1999) and the transcriptional profiling of the erythrocytic asexual stages of the parasite (Otto, et al, 2010).

In 2005 a large-scale analysis of about 3000 protein interactions has been performed by yeast-two-hybrid in *P. falciparum* (LaCount, et al, 2005)

allowing the reconstruction of the first PPI network. After that, other PPI networks have been proposed as reviewed in (Ramaprasad, et al, 2012). However, these PPI networks provide information only on a limited set of *P. falciparum* proteins. In 2006 a novel method based on the integration of "-omics" data from diverse source was adopted to derive a global PPI network (Date, et al, 2006; Hu, et al, 2010)

2 – AIM OF THE PROJECT

Antimalarial drug discovery and vaccine design have largely focused on compounds against blood stages of the parasite (Guiguemde, et al, 2012). The centrality of the blood stages has meant that extensive research efforts persist on specific biological processes commonly characterized by a manifold complex of membranes, such as erythrocyte invasion (Cowman & Crabb, 2006) and egress (Blackman, 2008). Furthermore, it has been shown that cell membranes are complex and inhomogeneously structures that perform and coordinate cell function by segregating or aggregating lipids and proteins. Despite of the complexity and relevance of the membrane system by which *Plasmodium* interacts with, enters and resides into the host cell, this area is relatively overlooked (Zuccala, et al, 2011).

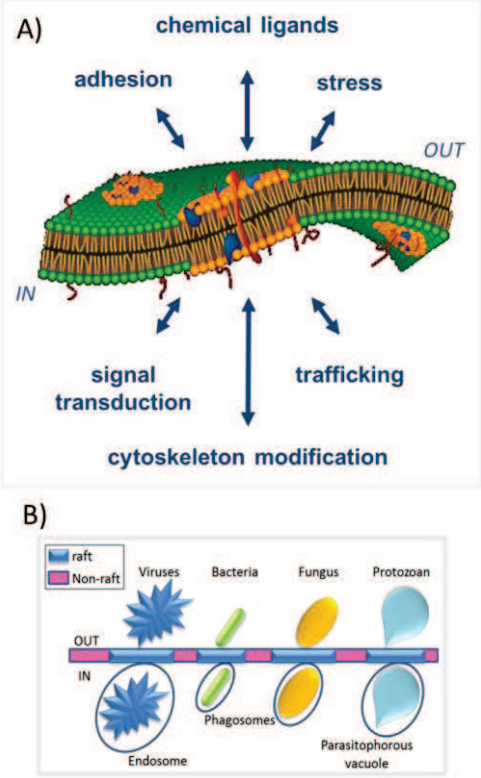


Figure 2. Lipid rafts structure and functions. A) Lipid rafts (in orange) distinct from the bulk membrane (green) are cholesterol-rich membrane microdomains that organize outside-in and inside-out responses to a wide range of stimuli; B) several pathogens target lipid rafts to enter the host cell.

Lipid rafts are cholesterol and sphingolipid-rich membrane compartments implicated in various cell processes by forming platforms that promote dynamic lipid-lipid, lipid-

protein, PPIs (Allen, et al, 2007; Lingwood & Simons, 2010) (panel A of figure 2).

In *Plasmodium*, as it occurs in a large set of other pathogens (panel B of figure 2), these membrane microdomains play a crucial role in host-cell invasion and other key processes like trafficking and signaling (Vieira, et al, 2010; Murphy, et al, 2006; Koshino, et al, 2009; Di Girolamo, et al, 2008). For these reasons, "raftophilic" proteins that preferentially and temporarily partition into the lipid raft may represent suitable potential drug targets (Rajendran, et al, 2010).

In this work, we propose a systems biology approach to study the dynamics of the highly specialized membrane microdomain also referred to as lipid rafts in *P. falciparum*.

3 – EXPERIMENTAL DESIGN

The peculiar richness of lipid rafts in cholesterol and sphingolipids is exploited to separate these microdomains from the rest of the membranes by sucrose density gradient. In the host laboratory, this method was applied to purify lipid rafts from *P. falciparum* young trophozoites, early schizonts, late schizonts, and

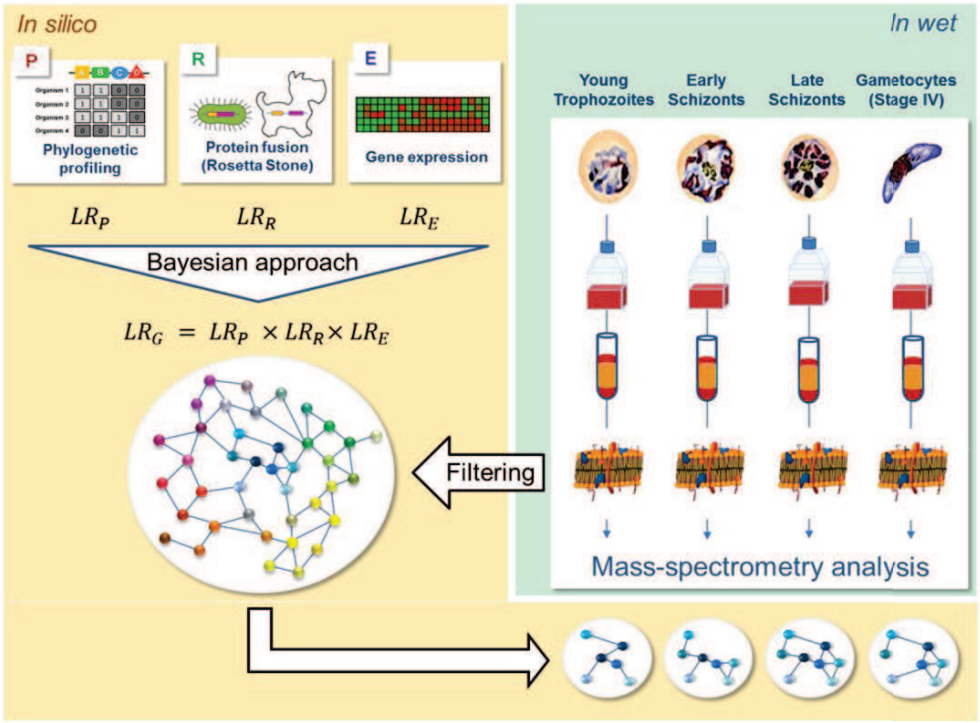


Figure 3. Experimental design. *In silico* approaches (evidences in light orange) are applied to predict PPIs from different data sources. A probability value representing the likelihood that an interaction occurs, is calculated from each dataset (LR_P from phylogenetic profiles, LR_R from rosetta stone data and LR_E from gene expression data). Then, a Bayesian approach is applied multiplying the probability values coming from each single dataset and calculating the probability that an interaction is likely occurring in the global interactome (LR_G). The *in wet* approach (evidenced in light green) was applied to purify the lipid rafts from different stages of *P. falciparum* erythrocytic life cycle and to analyze them by mass spectrometry. In the "filtering" procedure, the proteins identified as populating the lipid rafts at each stage are used to derive the corresponding stage-specific network.

mature gametocyte. Corresponding stage-specific proteomes, obtained by mass spectrometry, were available in public database (right side of figure 3).

My PhD thesis focused on a comprehensive analysis of PPIs taking place in lipid rafts of asexual and sexual *P. falciparum* blood stages. To do this, we derived stage-specific interactomes and studied their architecture, functional organization and dynamics.

We took advantage of the great amount of available "-omics" data to construct phylogenetic profiles, adopting a novel method to measure their similarity, to derive a novel set of "rosetta stone proteins" and a gene expression profiling of both asexual and sexual stages (left side of figure 3). We then predicted a global PPI network of *P. falciparum* applying the Bayesian approach proposed by Jansen (Jansen, et al, 2003).

Stage-specific proteomes were then used to filter the global PPI network and thus to derive the corresponding stage-specific interactomes of lipid rafts.

4 - RESULTS

4.1 Phylogenetic profile: the reassessment of the method.

The increasing number of genomic data makes phylogenetic profiling one of the most promising computational methods to predict PPIs. As mentioned in the Introduction, this method is based on the alignments between proteins from the genome under examination and proteins from selected genomes in a reference set, and on the comparison between phylogenetic profiles to detect possible co-evolving and thus interacting proteins. The two crucial points of this approach concern the selection criteria to include genomes in the reference set and the choice of the metric to measure the profile similarity.

Phylogenetic profiles of *P. falciparum* were constructed by Date and collaborators (Date, et al, 2006) using a reference set based on a limited number of genomes available at that time. Furthermore, to score the profile similarity they used the mutual information, whose limits in predicting positive correlations between arrays of real values have been recently recognized (Zhang, et al, 2015).

To overcome these limitations, in this work we proposed a novel procedure to construct phylogenetic profiles. We performed a re-assessment of the method by focusing first, on the strategy for the selection of genomes to be included in the reference set and then, on the measure used to score the profile similarity.

The number of genomes in a reference set should be sufficiently high to ensure a robust statistics, however, to avoid any kind of bias not redundant genomes should be included. A good balance between these two requirements would guarantee a robust statistics and hence a good predictive performances of the method. To do this, we developed a strategy of genome selection, based on the evolutionary relationships between species and the quality of the genome sequences. We started considering the 1133 organisms collected in the eggNOG database (http://eggnog.embl.de/version_3.0/; version 3.0) that are manually selected following several criteria, like genome coverage, status of gene annotation and gene completeness (Ciccarelli, et al, 2006; Powell, et al, 2012). According to these criteria, genomes in the EggNOG database are classified in two categories: core and peripheral. The "core" set includes only "high quality" genomes, whereas the "peripheral" set includes genomes sequenced more recently, that are not still completely validated. On the basis of this classification and on the inspection of the evolutionary relationships in the Tree of Life, constructed using genomes in eggNOG, we derived four genome reference sets of different size and genome composition (Table I).

The first reference set (RS1) was derived by excluding from the 1133 organisms all "peripheral" genomes of strains of the same species. A second reference set (RS2) was obtained excluding the eukaryotic genomes with a "peripheral" attribute from RS1. To construct the third reference set (RS3), we progressively excluded "peripheral" prokaryotic genomes till halving the size of RS1. The last reference set (RS4) was obtained from RS3 by visual inspection of the Tree of Life, excluding close phylogenetically related eukaryotic genomes to reduce possible redundancy.

Table I: characteristics of the reference sets.

	Total	*Prokaryotes*	*Eukaryotes*	*Archea*	*Ratio*
Reference set 1 (RS1)	774	594	120	61	5:1:0.5
Reference set 2 (RS2)	699	594	45	61	13:1:1.5
Reference set 3 (RS3)	336	230	45	61	5:1:1.5
Reference set 4 (RS4)	309	230	18	61	13:1:3.5

In order to assess the influence of both size and genome composition of the four reference sets on the predictive performance of the method, we constructed four phylogenetic profile data sets for *Saccharomyces cerevisiae* e *Escherichia coli* reference genomes using each of the four reference sets .

To align the *S. cerevisiae* and *E. coli* protein sequences against the proteins of the reference sets we used the Smith-Watermann alignment algorithm (Smith T.F. & Watermann M.S., 1981), and, as proposed by Date et al. (Date, et al, 2003), we constructed phylogenetic profiles as arrays of probability values obtained by the E-values, according to:

$$P = -1/\log_{10}(E) \qquad [1]$$

For E-values lower than 10^{-1}, the probability value is set to 0.

To compare phylogenetic profiles and hence to predict potential PPIs, we used a novel correlation metrics, the distance correlation (dCor). DCor was proposed by Szekely in 2007 as an extension of the Pearson's correlation, but at

difference with the linear coefficient, it measures any dependence between two variables X and Y. DCor satisfies the properties:

a) $0 \leq \text{dCor}(X,Y) \leq 1$

b) $\text{dCor}(X,Y) = 0$; if and only if X and Y are independent

c) $\text{dCor}(X,Y) = 1$; if $Y = \alpha + \beta(X)$ where α and β are scalar

d) $\text{dCor}(X,Y) \neq 0$; if $Y \approx f(X)$

To perform a comparison between dCor and Mutual Information (MI), usually used to measure phylogenetic profile similarity, we carried out an assessment of the method utilizing both dCor and MI.
The mutual information between profiles X and Y was calculated as:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \qquad [2]$$

where $H(X) = -\sum p(x) \ln p(x)$ is the summation of the marginal entropies, calculated over the intervals of probability distribution $p(x)$, and $H(X,Y) = -\sum\sum p(x,y) \ln p(x,y)$ represents the summation of the relative entropies of the joint probability distribution $p(x,y)$. Phylogenetic profiles are continuous variables and to apply MI they have to be discretized by binning the real values into intervals, and then deriving the marginal and the joint probabilities. MI was calculated by using the *mutualInfo* function (Joe, 1989) available in *bioDist* package (Ding, et al, 2010) for R environment (R Development Core Team, 2012) after binning the data into 0.1 intervals.

While the MI calculation was performed by applying already available scripts, in the case of dCor we needed to develop a novel software that parallelizes the several steps of the computation making it possible the application of this statistics to large genomic data like phylogenetic profiles.

The strategy that we adopted is schematically shown in figure 4. The first part of the procedure (Process 1) represents the phase of data preparation and includes three steps. The step 1 consists in the construction of the matrix with

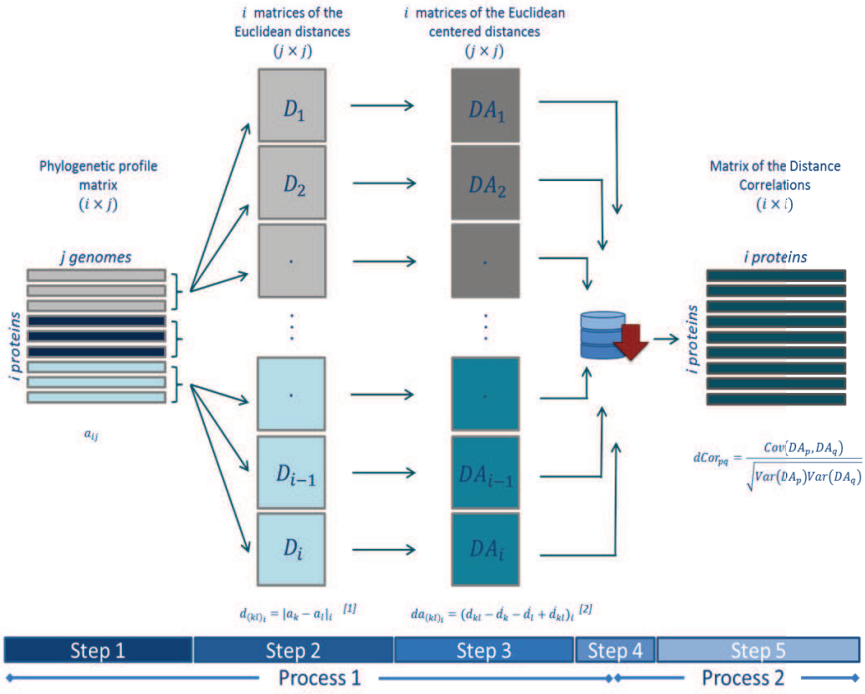


Figure 4. Workflow. The calculation of the distance correlation has been divided into two processes. By the Process 1, a "split-apply" strategy is adopted (Step 1): the phylogenetic profile matrix is divided in subsets of profiles that are contemporary processed for calculations (groups of bars in different colors). From each single phylogenetic profile is calculated the matrix of the Euclidean distances (Step 2). Each element of the Euclidean distance matrix is calculated as the absolute difference between the *k* and *l* element of the *i*-th phylogenetic profile. The Euclidean distance matrixes are represented by light grey or light blue squares in the figure. Then each element of the Euclidean distances matrix is "centered" by following the formula [2] as reported in the figure at Step 3, obtaining the matrix of the centered distances represented by dark gray and dark blue squares. Then, each matrix of the centered distances is saved as a binary file (Step 4) creating a repository. The final step (Step 5) in Process 2, represents the calculation of the distance correlation from each pair of matrixes of the centered distances read from the binary files in the database.

elements a_{ij} containing the phylogenetic profiles of i proteins of a given organism, constructed using a reference set of size j genomes; starting from this data, the i Euclidean j x j distance matrices D with elements $d_{(kl)i} = \left| a_k - a_l \right|_i$ between the j elements of the i-th phylogenetic profile (with k and l from 1 to j) are calculated (step2); the step3 regards the calculation of the centered Euclidean distances DA of elements $da_{kl} = d_{kl} - \overline{d}_k - \overline{d}_l + \overline{d}_{kl}$ where

$\overline{d}_k = \frac{1}{n}\sum_{k=1}^{n} d_{kl}$ is the average calculated on the rows of the distance matrix;

$\overline{d}_l = \frac{1}{n}\sum_{l=1}^{n} d_{kl}$ is the average calculated on the columns of the distance matrix;

$\overline{d}_{kl} = \frac{1}{n^2}\sum_{k,l=1}^{n} d_{kl}$ is the average calculated on all the elements of the distance matrix;

with $k = l = 1,...., n = 1,..., j$.

This is the crucial point of the procedure, in fact, the calculation of centered Euclidean matrices depends on sample size, according to a polynomial law. This hampers the application of the method to a large data set on a standard workstation. To reduce the algorithm complexity we applied a "split-apply-combine" strategy developing a code that subdivides and parallelizes this step. Furthermore, to reduce RAM requirements, matrices resulting from step 3 are stored as binary files in a repository (step4).

To perform the second part of the procedure (Process 2), a second code was developed to read pairs of binary files *(p, q)* and to calculate the corresponding

distance correlation $dCor_{pq} = \frac{Cov(DA_p, DA_q)}{\sqrt{Var(DA_p)Var(DA_q)}}$ where $Var(DA_p)$ and $Var(DA_q)$ are variances of matrices DA_p and DA_q and $Cov(DA_p, DA_q)$ is their covariance.

Each profile of the four sets of phylogentic profiles (constructed using the reference sets reported in table I) was compared with all the others in the same data set. This resulted in eight data sets of PPI predictions, for *S. cerevisiae* and *E. coli* with the corresponding dCor values and eight datasets with the corresponding MI scores.

In order to perform an assessment of the method, we derived two sets of gold standards both for *S. cerevisiae* and *E.coli*. The first one (GS-fun) was derived from the metabolic pathways in KEGG database (Kanehisa & Goto, 2000), and includes, as True Positives (TPs), pairs of functionally related proteins (TP-fun) belonging to the same pathway. In the second set of gold standards (GS-phy), TPs are pairs of physically interacting proteins (TP-phy) derived from the STRING database (Franceschini, et al, 2013) considering protein pairs with exclusively experimental evidence of a physical interaction. To compose True Negatives sets (TN) of both functional (TN-fun) and physical (TN-phy) interactions, we developed a graph-based algorithm to explore data in KEGG or in STRING databases (figure 5). Data were represented as networks where interconnected nodes are, in one case, proteins of the same pathway, while in the other proteins of the same complex. The algorithm automatically explores the networks to identify protein pairs whose shortest path between their sub-graph was higher or equal to five (TN).

To assess the performance of our predictions, a ten-fold cross-validation procedure was applied. Each dataset of protein pairs was randomly sub-divided in independent subsets of equal size (ten subsets of 3465677 protein pairs each for yeast data and ten subsets of 2011522 protein pairs each for *E. coli* data) and the corresponding specificity and sensitivity values were calculated according to:

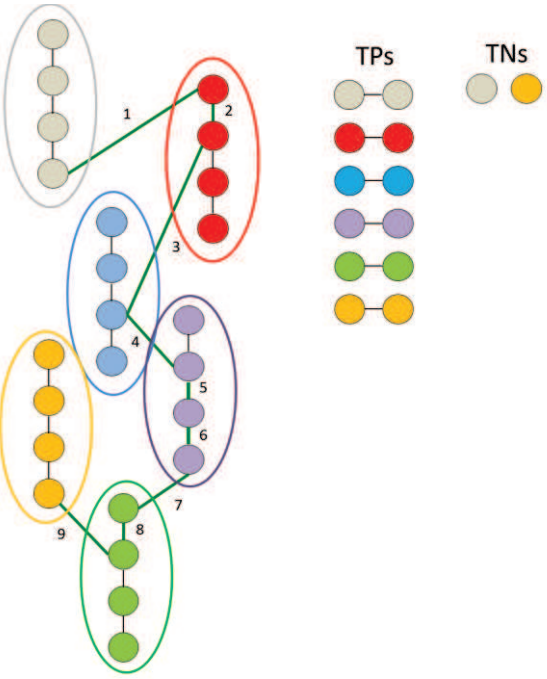


Figure 5. Gold standard construction. A graph based algorithm was developed to construct gold standards. Data from KEGG and STRING databases are represented as graphs, where proteins in the same metabolic pathway (KEGG) or in the same complex (STRING) are, represented as nodes of the same color in the figure. True Positives are derived as protein pairs in the same pathway or in the same complex. To find True Negatives shortest paths between nodes in the graphs are calculated (green thick line) and those pairs, which shortest path is equal or higher than 5 are selected as TNs.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

where *TP, FP, TN, FN* are the number of true and false positives and negatives obtained by comparing protein pairs (at each interval of dCor and MI values binning them at 0.1 intervals) with gold standards. As a measure of the prediction accuracy, the Area Under the ROC Curve (AUC) was calculated according to the Gini's formula:

$$AUC = \frac{1}{2}\sum_i ((X_i - X_{i-1})(Y_i + Y_{i+1})) \qquad [3]$$

where $X_i = 1 - Specificity$ and $Y_i = Sensitivity$ at the i-th bin of dCor and MI.

Results, obtained for each assessment performed on phylogenetic profiles constructed using each reference set (RS), are shown as scatter dot plots in figure 6 where means and standard deviations are indicated as horizontal and vertical bars respectively. In all cases, the predictive performance of

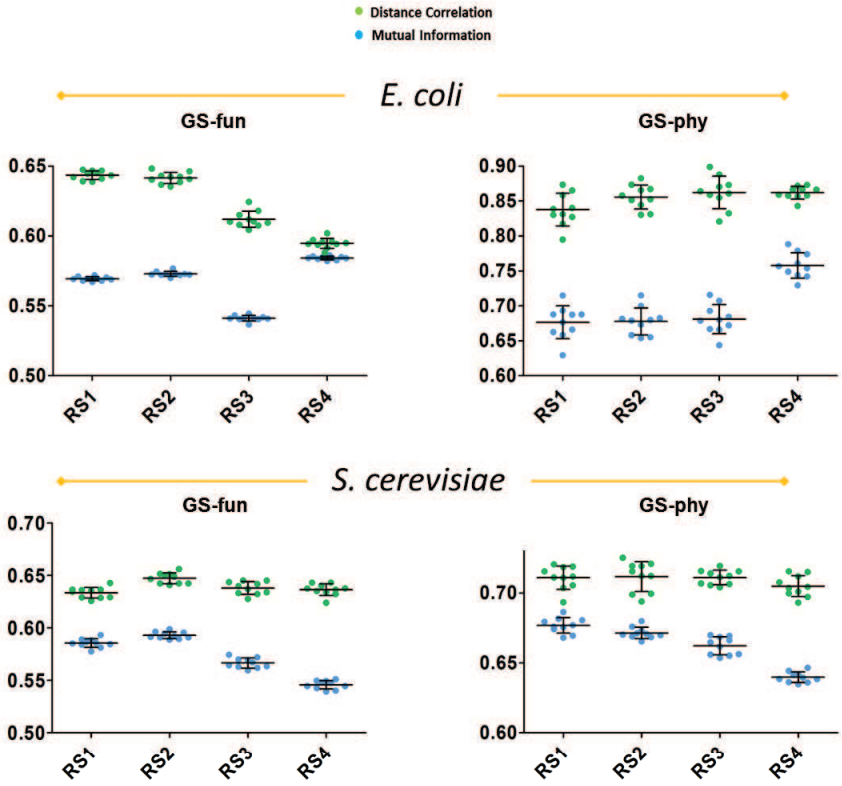


Figure 6. Assessment results. A ten-fold cross-validation was performed to assess predictive performances by using dCor (green dots) and MI (blue dots) to measure phylogenetic profile similarity. Each dot represents the prediction accuracy calculated as the Area Under the ROC Curve (AUC). The phylogenetic profiles of *E.coli* and *S.cerevisiae* were constructed comparing protein sequences with those from genomes in four reference sets (RS1, RS2, RS3 and RS4). In order to assess the performance of our method in predicting both functional and physical interactions, gold standards were derived from KEGG database as "functionally related protein" (GS-fun) and from STRING database, as "physically interacting proteins" (GS-phy; considering only the protein pairs for which evidence of a physical interaction was available).

phylogenetic profiling using dCor (green dots) outperforms that one obtained using MI (blue dots). In particular, in the functional interaction predictions, AUC mean values range from 0.60 to 0.64 for dCor and from 0.53 to 0.58 for MI, in the case of *E. coli* and from 0.63 to 0.65 for dCor and from 0.55 to 0.59 for MI, in the case of *S. cerevisiae*. Even better results were obtained in predicting physical interactions, AUC mean values ranging between 0.84 and 0.86 for dCor and between 0.68 and 0.76 for MI for *E.coli*, and between 0.70 and 0.71 for dCor and between 0.64 and 0.68 for MI in the case of *S. cerevisiae*. Considering overall data, the best predictive performances were achieved with the reference sets RS1 and RS2 that contain the highest number of genomes (774 and 669 respectively). However, RS2 appears to outperform RS1 in predicting physical interactions in *E.coli* and the functional ones in yeast.

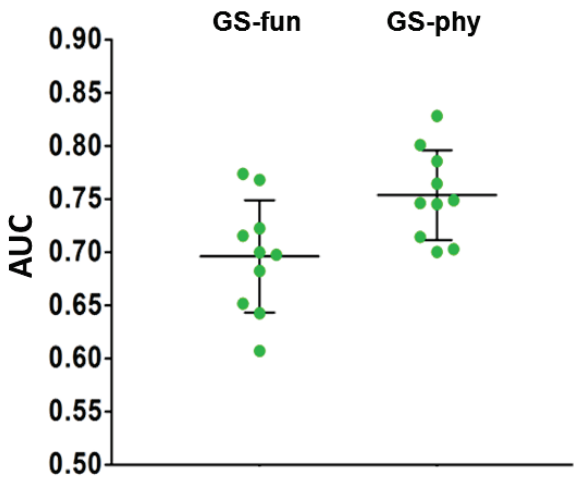


Figure 7. *P. falciparum* assessment results. The phylogenetic profiles of *P. falciparum* were constructed comparing their protein sequences with protein sequences from genomes in the reference sets RS2. The distance correlation was used to score profile similarity. Predictions were assessed by the comparison with the gold standards derived from KEGG (GS-fun) and the gold standards derived from STRING (GS-phy).

For this reason, we proceeded with the construction of a novel set of phylogenetic profiles for *P. falciparum* using the reference set RS2 and the distance correlation, as profile similarity measures. Results obtained in the assessment procedure are

shown in figure 7 as scatter dot plots. As in the case of *E. coli* and *S. cerevisiae* the prediction of physically interactions (AUC mean value = 0.75) gives better results compared with those of the functional interactions (AUC mean value= to 0.69).

4.1.a A gold standard-independent validation: the case of the apicoplast proteins of *P. falciparum*.

Organelle proteins in *P. falciparum* are of particular interest as candidates for drug development in that they are potentially specific for the parasite and absent in the host cells. In this context, the apicoplast proteins appear of interest because of their bacterial origin, and because they are involved in biological processes essential for the parasite life. As shown in the figure 8, the apicoplast originates from a secondary endosymbiosis event occurred when an eukaryotic red alga engulfed a cyanobacterium and then was subsequently engulfed by a eukaryotic cell (McFadden, et al, 1996). During the evolution, several gene transfer events occurred from the cyanobacterium to the red alga genome and then from that to the nuclear genome of the eukaryotic host cell. This made the apicomplexan genomes a sort of mosaic of genes from different origins. Nowadays the apicoplast genome encodes a small number of proteins involved in the metabolic pathways that take place in the organelle (Wilson, et al, 1996), and the majority of the proteins (of bacterial and plant origin) that are necessary to complete its functionality are encoded by the nuclear genome and targeted to the apicoplast by localization signals (Martin & Herrmann, 1998; Foth, et al, 2003).

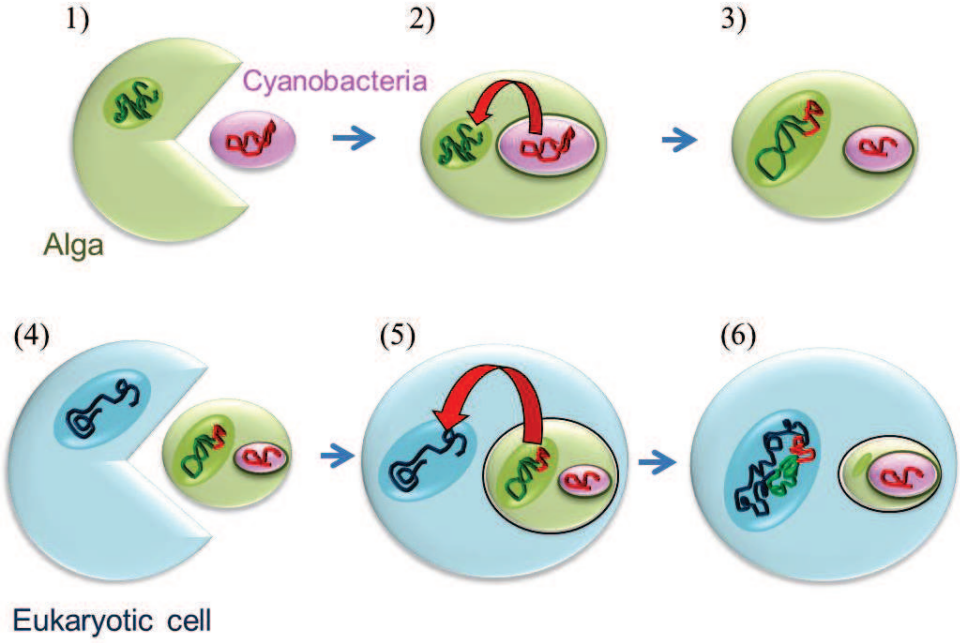


Figure 8. Evolutionary origin of the apicoplast. The apicoplast derived from a double event of endosymbiosis. In the first endosymbiosis a cyanobacteria is engulfed by an alga (1) and subsequent gene transfer events happened from the cyanobacterium to the alga genome (2-3). A second endosymbiotic event occurred when the alga was engulfed by an eukaryotic cell (4) determining a second series of gene transfers (red arrow) from the alga's nucleus to the eukaryotic one (5-6).

Methods have been developed to predict apicoplast proteins on the basis of target peptide sequence (PATS, PlasmoAP and ApicoAP: Zuegge, et al, 2001, Foth, et al, 2003, Cilingir, et al, 2012), however, attempts to define a consensus for this localization signal failed. Furthermore the low number of experimentally validated apicoplast proteins makes problematic the development of algorithms based on the usage of a training set.

In this work, we utilized phylogenetic profiles of *P. falciparum* to investigate evolutionary relationships of known apicoplast proteins, and to evaluate the possibility to exploit this method in the prediction of novel ones.

To do this, we first established a threshold of the dCor on the basis of the TP/FP ratio *versus* the genome coverage. We choose a value that guarantees a good enough balance between the coverage of the parasite proteome, and thus a sufficient number of predicted interactions, and an acceptable percentage of possible FPs. We choose the threshold dCor > 0.6, that corresponds to 1928 proteins involved in 46712 predicted interactions with a ratio of TP/FP equal to 32.9 and hence a percentage of 3% of possible FPs (figure 9).

The resulting data set of protein pairs was represented as a network by Cytoscape (version 2.8.3; Shannon, et al, 2003), in which (figure 10) each protein is shown as a node (dots in figure) and evolutionary associations between proteins, as established by their dCor value, as edges.

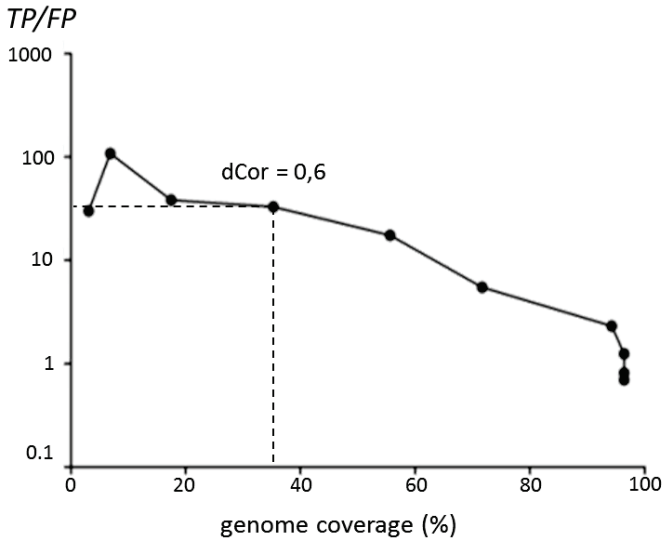


Figure 9. Assessment of *P. falciparum* phylogenetic profiling. At each dCor value, the TP/FP ratio and the percentage of proteins encoded by the *P. falciparum* genome (genome coverage) are calculated. We establishe a treshold of dCor= 0.6, that corresponds to a genome coverage of 35.3% and a TP/FP of 32.87.

Then, to confirm the predicted associations, we exploited information in the ApiLoc Database (http://apiloc.biochem.unimelb.edu.au/apiloc/apiloc) regarding experimental evidence on sub-cellular localization of *P. falciparum* proteins. 23 out of 40 apicoplast proteins selected from the ApiLoc Database

were found to map on the network of phylogenetic profiles. As shown in figure 10, known apicoplast proteins localize in two separated regions of the network. 15 proteins reside in a central area of the main network (figure 10, green nodes), while the remaining 8 proteins (figure 10, red nodes) remain confined in a small network that has no connections with the main one. This result strongly suggests that the small network includes proteins encoded by genes originally derived from the cyanobacterium genome and hence evolutionary unrelated with the rest of the proteome. On the other hand, apicoplast proteins encoded by genes with "plant origin" are not sparse, but they localize to a limited region of the main network, reflecting their common evolution.

Then, we focused on the subnetwork containing proteins of bacterial origin. This contains 60 proteins, which are involved in pathways known to take place in the apicoplast such as the type II fatty acid synthesis (FASII) or the iron-sulphur cluster synthesis (Lim, et al, 2010). In particular, we found that three subunits of the Pyruvate Dehydrogenase (PDH) complex in ApiLoc, map in this subnetwork. The PDH complex, that is involved in essential *Plasmodium* processes (Pei, et al, 2010), is formed by four subunits (E1α, E1β, E2 and E3). No information is available on the sub-cellular localization of the subunit E1β (PF14_0441) and no bioinformatics tool predicted it as an apicoplast protein. As shown in the figure (magnification in figure 10), we found that this protein is linked to the other three subunits of the PDH complex in the subnetwork. This result represents a further validation of the phylogenetic profiling that we constructed for *P. falciparum* and opens new perspectives on its potential applications in developing a novel class of predictors.

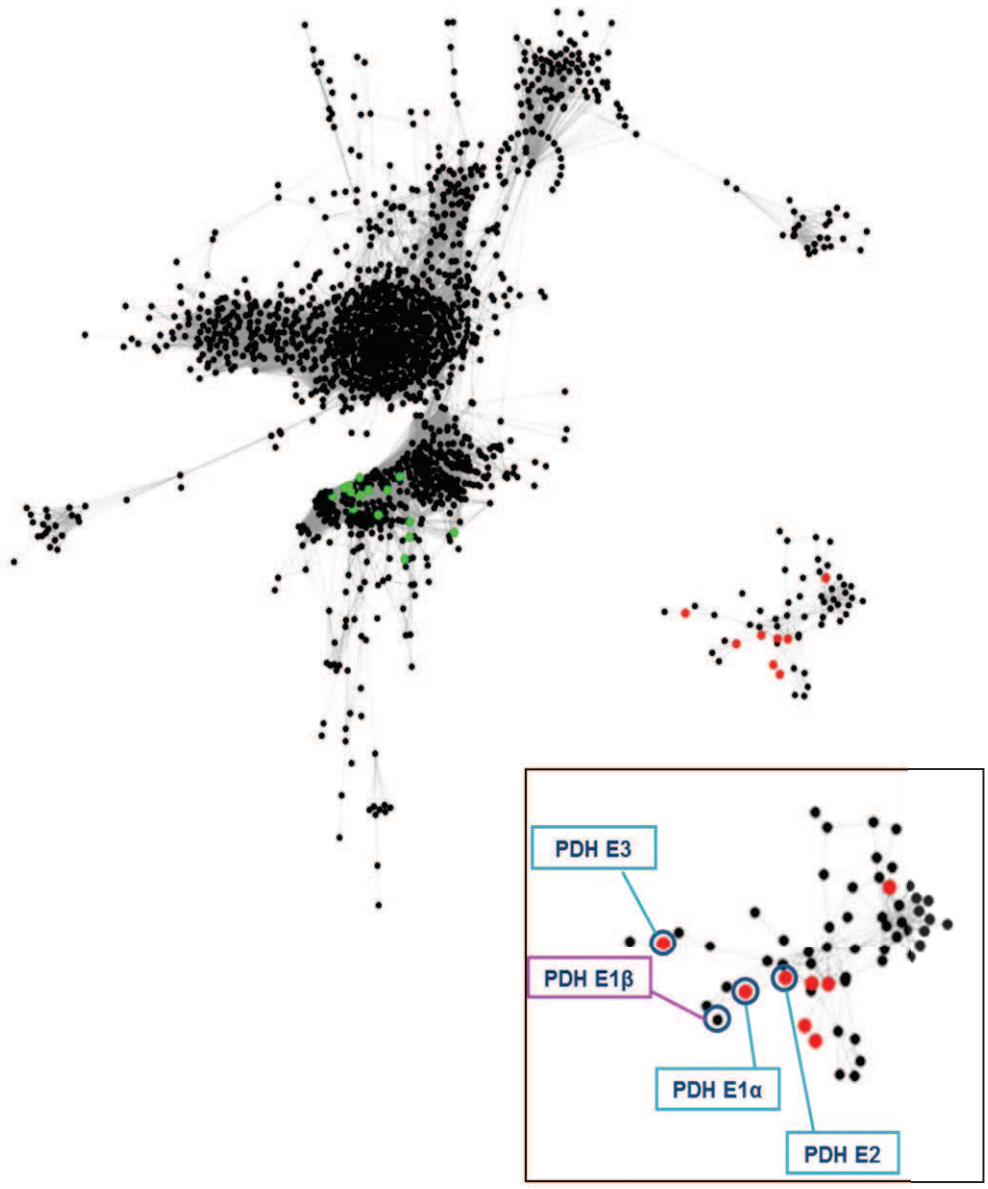


Figure 10. *P. falciparum* phylogenetic profiles relationships are represented as a network. We selected a subset of protein pairs of *P. falciparum* with a phylogenetic profile similarity higher than 0.6. The proteins and their interactions were visualized by Cytoscape (v2.8.3) as a network. Experimentally validated apicoplast targeted proteins are shown as colored nodes. Green nodes are apicoplast proteins with an hypothetical red alga origin, while red ones represent proteins encoded by genes of bacterial origin. In the inset the "bacterial-related" network is highlighted. The proteins belonging to the PDH complex are in blue circle.

4.2 Rosetta stone data of *P. falciparum*.

Rosetta stone is a method to predict PPIs, based on the detection of "rosetta stone proteins" in genomes of a reference set (Enright, et al, 1999; Marcotte, et al, 2002). The basic idea is that two not-paralog proteins are predicted to be functionally related if they are similar to two different not-overlapping domains of the same protein in the reference set (figure 11).

We utilized the reference set RS1 (774 genomes; $4x10^6$ protein sequences) to construct novel rosetta stone fusion data of *P. falciparum* proteins.

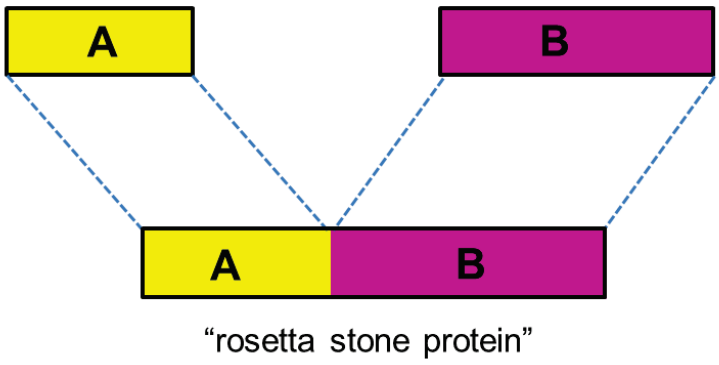


Figure 11. Rosetta stone method. The method predicts the interaction between two proteins A and B (yellow and purple bars) if they have, in another organism, the orthologs fused in a multi-domain protein ("rosetta stone protein"). The orthologs of the target proteins are detected by Smith-Watermann sequence alignment. Matches are considered only if the resulting E- value is lower than 10^{-6} and the sequence identity is higher than 30%.

The protein sequences of *P. falciparum* were aligned against the proteins derived from the genomes in the reference set by the Smith-Watermann alignment algorithm available in the FASTA package (version 36, Pearson & Lipman, 1988). A match between a parasite protein and a "rosetta-stone protein" was considered if E-value of the alignment is lower than 10^{-6} and sequence identity higher than 30%. *P. falciparum* paralogs were excluded from the analysis, since proteins with a percentage of identity higher than 50% were discarded.

In order to establish the likelihood of the predicted functional interaction, we calculated the probability that a fusion event occurred by chance. As proposed in (Marcotte and Marcotte, 2002), we used the hypergeometric distribution to calculate the probability of observing k fusion events between two proteins that occur n and m times in a database of size N genes, according to:

$$p(k|n,m,N) = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}} = \frac{n!(N-n)!\,m!(N-m)!}{(n-k)!\,k!(m-k)!(N-n-m+k)!\,N!} \quad [4]$$

To calculate large factorials, the logarithm of the probability using Sterling's approximation was calculated. Finally, the probability to observe k or more fusion events by chance is equal to:

$$p(fusions \geq k|n,m,N) = 1 - \sum_{i=0}^{k-1} p(i|n,m,N) \quad [5]$$

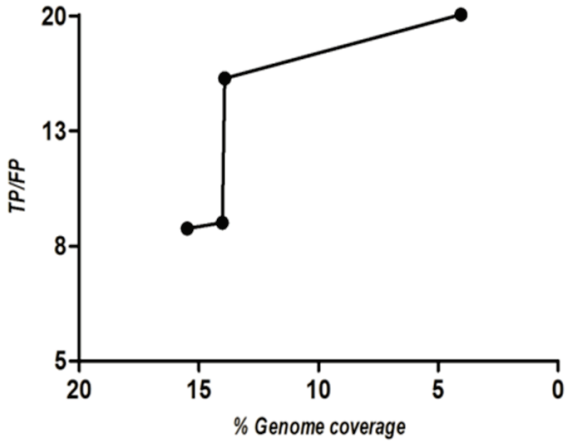


Figure 12. Rosetta stone assessment. TP/FP ratio and genome coverage calculated for each of rosetta stone probability scores.

The lower the probability value, the more significant the inferred interaction is.

846 *P. falciparum* proteins, involved in 620 interactions were predicted as functionally related by this procedure. We evaluated the predictive performance of the results comparing the obtained

data with the functional gold standards (GS-fun) already used in phylogenetic profiling. Figure 12 shows the outcome of the assessment procedure. Although the predicted interactions regard a limited number of proteins, our data show a good predictive performance (the ratio TP/FP ranges between 8.5 and 20.0) that is higher than the previously published (TP/FP < 10) (Date, et al, 2006).

4.3 Expression profiling of *P. falciparum* blood stages.

The prediction of PPIs from expression data is based on the observation that proteins in the same complex, or involved in the same pathway, are often encoded by co-expressed genes. Many transcriptomic data are now available for *P. falciparum,* that can be utilized to determine gene co-expression. We focused on microarray-based expression data for which both asexual and sexual blood stages (gametocytes) have been analyzed (Le Roch, et al, 2003; Young, et al, 2005).

Transcriptomic data are accessible for early and late rings, early and late trophozoites, early and late schizonts, merozoites and gametocytes (Le Roch, et al, 2003). For asexual stages two data sets are available corresponding to two independent cultures synchronized using different methods (sorbitol and temperature). Two additional gene-expression data sets are also available for sexual stages from samples collected daily from two different *P. falciparum* lines (NF54 and 3D7). The first data set consists of a time course of NF54 gametocytes (from day 1 to day 13) while the second of a time course of 3D7 gametocytes (days 1, 2, 3, 6, 8 and 12) (Young, et al, 2005).

Starting from these data, we constructed different profiles as summarized in table II.

Table II: length and composition of the expression profiles

	Total number of time points	*Asexual stages*		*Sexual stages*	
		Sorbitol synchronization	*Temperature synchronization*	*3D7 clone*	*NF54 clone*
Profile 1	26	7 [a]	7 [a]	6 [b]	6 [b]
Profile 2	14	7 [a]	7 [a]	-	-
Profile 3	12	-	-	6 [b]	6 [b]
Profile 4	20	7 [a]	7 [a]	6 [b]	-
Profile 5	15	7 [a]	7 [a]	1 [a]	-

[a] from Le Roch, et al, 2003; [b] from Young et al, 2005.

Raw expression data were downloaded from PlasmoDB Database (http://plasmodb.org). Each expression profile value was then log2 transformed and normalized according to:

$$x_{normalized} = \log_2 x_i / \bar{x}_{profile} \qquad [6]$$

where $x_{normalizea}$ is the normalized value of expression calculated for the *i*-th *x* element of the expression profile and where $\bar{x}_{profile}$ is the mean calculated along the profile.

For each pair of profiles the corresponding Pearson's correlation coefficient was calculated to measure profile similarity (the *cor* function (Becker, et al, 1988) in R environment). We assessed the results using GS-fun gold standard obtained from KEGG database. The Positive Predictive Value (PPV) was calculated according to

$$PPV = \frac{TP}{TP + FP} \qquad [7]$$

where *TP* and *FP* were calculated at each Pearson's correlation value. The results of the assessments are reported in figure 13. We found that the best PPVs were obtained using profiles 2 and 5 while the predictive performances using the other profiles was significantly lower.

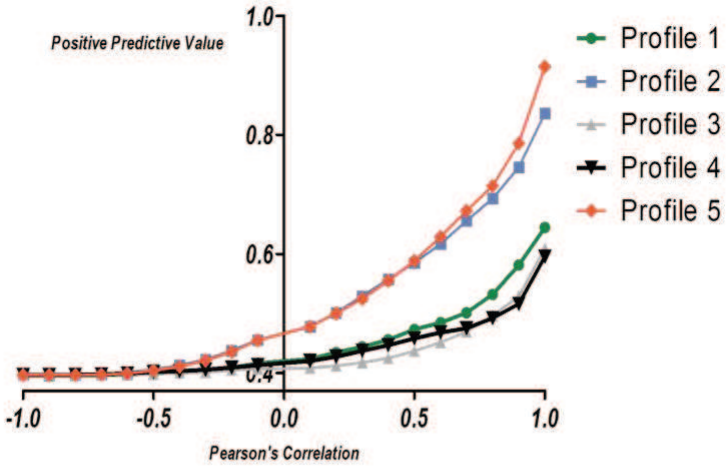


Figure 13. Gene expression assessment. The Positive Predictive Value was calculated for each value of the Pearson's coefficient obtained by comparing gene expression profiles in each data set (Profile 1-5).

4.4 Prediction of a global *P. falciparum* interactome.

We utilized phylogenetic profiles, rosetta stone proteins and gene expression data to predict a global PPI network for *P. falciparum.* We applied the Bayesian approach proposed by Jansen (Jansen, et al, 2003) to integrate the three data sets.

Protein pairs in each data set were grouped binning the corresponding value of Pearson's correlation, distance correlation and rosetta protein probability. Each group of protein pairs in each bin was compared with the gold standards of functional interactions (GS-fun) and the corresponding TP/FP ratio was calculated (Table III). This TP/FP value was then associated to each protein pair in the bin as the score *LR*.

Table III. TP/FP ratio calculated for each bin of the dCor, rosetta stone probability and Pearson's coefficient

Phylogenetic profile		Rosetta stone data		Expression data	
bin	TP/FP	bin	TP/FP	bin	TP/FP
0 - 0.1	0.696261	< 5	0	-0.9 – -0.8	0.65716
0.1 0.2	0.81191	5 – 6	8.75	-0.8 – -0.7	0.65724
0.2 - 0.3	1.24575	6 – 7	8.5	-0.7 – -0.6	0.65742
0.3 - 0.4	2.3144	7 – 8	15.5	-0.6 – -0.5	0.65942
0.4 - 0.5	5.47423	else	20	-0.5 – -0.4	0.66545
0.5 - 0.6	17.4959			-0.4 – -0.3	0.6781
0.6 - 0.7	32.8667			-0.3 – -0.2	0.69691
0.7 - 0.8	38.2857			-0.2 – - 0.1	0.72763
0.8 - 0.9	108			-0.1 – 0	0.7715
0.9 - 1	30			0 – 0.1	0.83451
				0.1 – 0.2	0.9149
				0.2 – 0.3	1.0003
				0.3 – 0.4	1.10305
				0.4 – 0.5	1.24752
				0.5 – 0.6	1.43028
				0.6 – 0.7	1.69389
				0.7 – 0.8	2.05295
				0.8 – 0.9	2.50138
				0.9 – 1	3.67123
				1	10.6897

A global score LR_G was obtained by multiplying the single ones derived from each data set,

$$LR_G = LR_P \times LR_R \times LR_E \qquad [8]$$

here LR_P, LR_R and LR_E are the scores from the phylogenetic profiles, rosetta stone proteins and expression data respectively.

An assessment of the global predictive performance was performed and compared with those derived from each single data set by using the gold standards GS-fun. Results are shown in figure 14 as a plot of the TP/FP for each

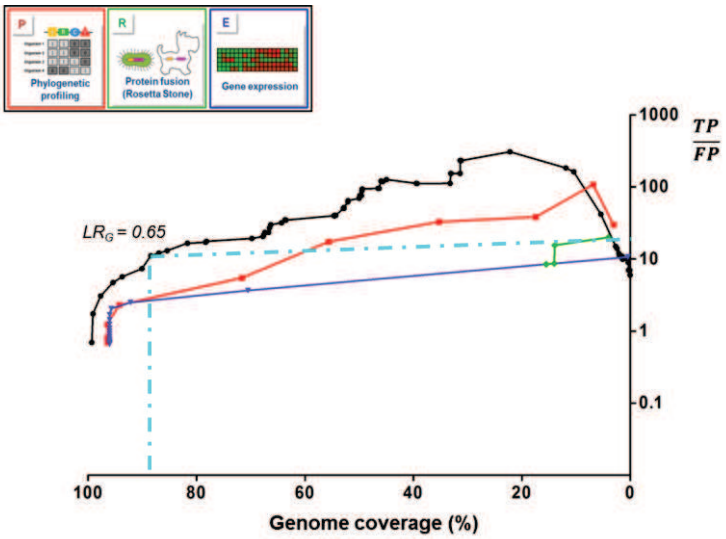


Figure 14. Assessment of the *P. falciparum* PPIs. The assessments of PPIs predicted by phylogenetic profiling (red line), rosetta stone method (green line), gene expression data (blue line) and by the Bayesian approach (black line) were performed. Accuracy (TP/FP) and genome coverage are calculated for each of dCor, rosetta stone probability, Pearson's coefficient and for the global score LR_G. A threshold was established for LR_G equal to 0,65 that corresponds (light blue line) to more than 88% of genome coverage and to a ratio TP/FP equal to 11.1.

LR_G value *versus* the genome coverage.

Our results show that the integration of data results in an improvement of the accuracy in predicting PPIs. The TP/FP ratios of the predictions obtained using integrated data sets is about ten-fold higher compared with the TP/FP ratios obtained using the single data sets. In order to derive a PPI network covering a significant portion of the proteins encoded by *P. falciparum* genome and ensuring a sufficiently low number of expected FPs, we choose a LR_G value equal to 0.6 (88% of genome coverage; 9% of FPs) as a threshold. The resulting PPI network includes 4833 proteins involved in 436115 interactions (figure 15).

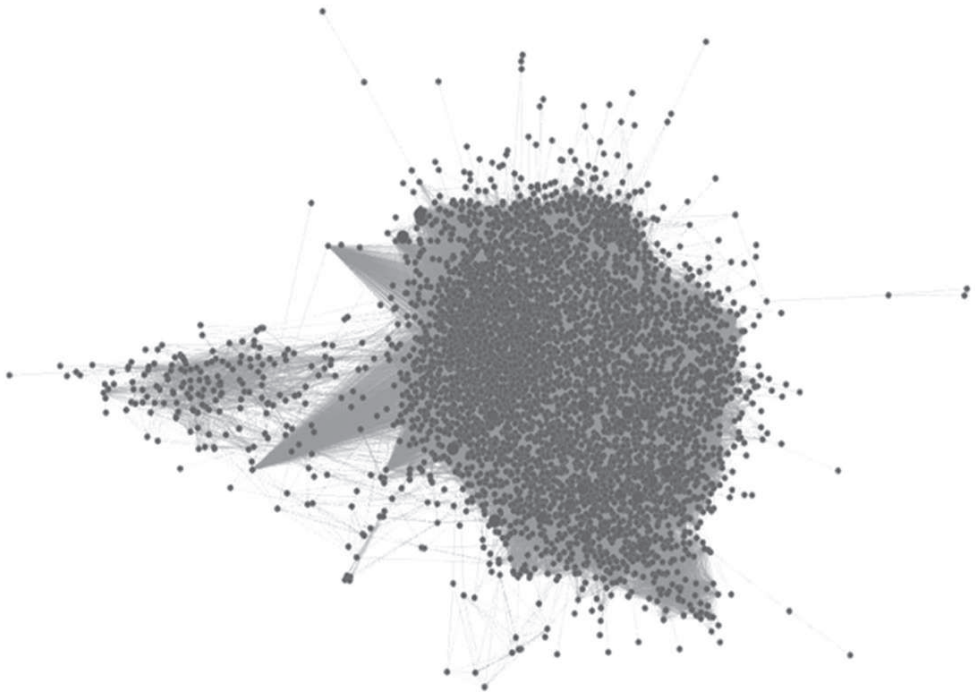


Figure 15. *P. falciparum* PPI network. Predicted PPI network is represented as a graph. Each node represents a protein while edges represent predicted interactions.

4.5 Topological analysis of *P. falciparum* global interactome.

Biological networks are characterized by specific topological properties that distinguish them from random. First, biological networks are scale-free (panel a of figure 16), that is they have few nodes with a high number of connections (high node degree) and many nodes with few connections (low node degree)

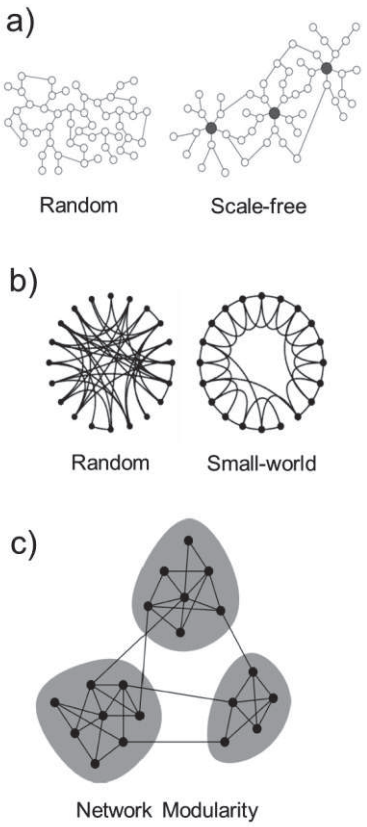


Figure 16. Topological properties of biological networks. Biological networks have a scale-free feature (panel a). In these networks few nodes have a high number of connections (full nodes), while the majority of them are poorly connected (empty nodes).
The small-world property is another feature of biological networks (panel b). The majority of nodes, even if not directly connected, can be reached from every other node by few steps.
Biological networks are modular (panel c). Groups of highly connected nodes (grey areas) are poorly connected.

(Uetz, et al, 2000; Ito, et al, 2001). This results in a node degree distribution that follows a power- law $y \approx x^{-\gamma}$ where γ ranges from 2 to 4 (Barabási & Albert, 1999). Second, biological networks have small-world property (panel b in figure 16), that is the majority of nodes, even if not directly connected, can be reached by a small number of paths (Watts & Strogatz, 1998). Finally,

biological networks are modular. They are formed by highly connected groups of nodes that are poorly connected to each other (panel c of figure 16) (Newman, 2006). Modularity of a network can be estimated by comparing the distribution of the average clustering coefficients with that one of a random network. The clustering coefficient is defined as the ratio between all possible and the actual edges between first neighbors of a given node. The distribution is obtained by calculating the average of the clustering coefficient of nodes with 1,2, 3,……n first neighbors.

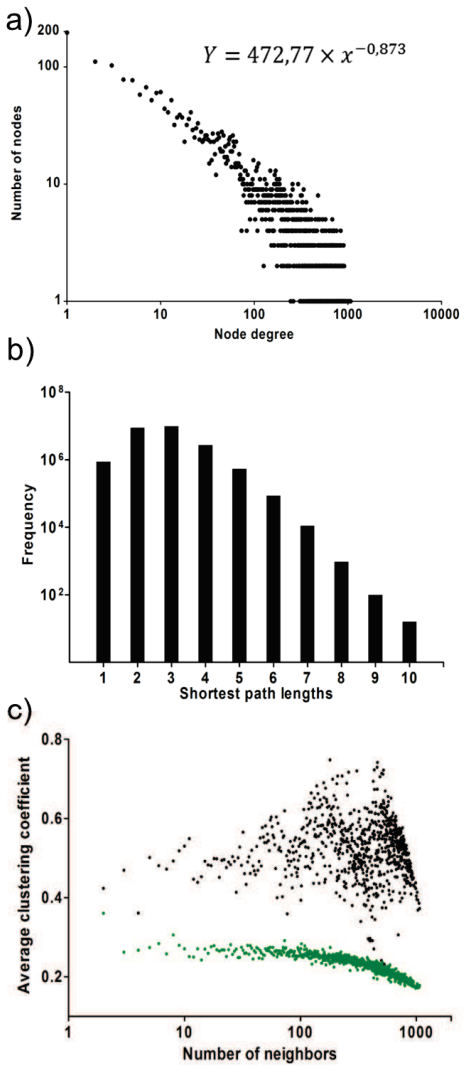


Figure 17. Features of *P. falciparum* interaction network. We analyzed the interactions above the threshold of 0.6 of likelihood score, calculating:
a) The distribution of node degree: each dot represents the number of nodes that have a specific number of connections (degree). The power law fitting with the data is reported on the top right of the panel. Note the log-scale on x and y axes.
b) The shortest path lengths distribution: bar graph of the distribution of the shortest paths length (Y-axis in log-scale).
c) The distribution of the average clustering coefficient: each black dot represents the average of the clustering coefficient calculated for all nodes with the same number of neighbors of *P. falciparum* network, while the green dots represent the same data calculated for a random network of equal size (X-axis in log-scale).

To analyze the PPI network obtained for *P. falciparum*, we used the

Network Analysis tool available in Cytoscape (version 2.8.3, Shannon, et al, 2003), and results are shown in figure 17. In panel a of figure 17, it is shown the node degree distribution that fits the power law $y = 472.77 \times x^{-0.873}$. We noticed that the γ value is lower than the one expected for biological networks. This means that a high number of nodes has high node degree in *P. falciparum* PPI network. This result supports previous observations (La Count, et al, 2005; Date, et al, 2006) indicating that *P. falciparum* interactome is more interconnected than those of other organisms. Panel b of figure 17 shows the distribution of the shortest path lengths, we observed that the majority of node pairs is characterized by shortest path lengths lower than 5 (96,56%), revealing the "small word" character of the *P. falciparum* interactome. In panel c of figure 17, the distribution of the average clustering coefficients are reported for both real (full dots) and randomized (green dots) *P. falciparum* interactome. The two distributions significantly differ. Average clustering coefficients obtained for the real interactome are consistently higher than those obtained for the randomized one (the global average clustering coefficient of the real interactome is equal to 0.493, while it is equal to 0.242 for the randomized one).

4.6 Stage-specific lipid raft interactomes during *P. falciparum* blood stage development.

The prediction of a global interactome does not take into account timing of expression and sub-cellular localization of the proteins involved. To overcome this problem, it can be filtered with proteomic data. The host lab performed a proteomic analysis of membrane microdomains (lipid rafts) isolated from

different *P. falciparum* blood stages. Lipid rafts, characterized in many different organisms, are involved in important cellular processes such as signaling, trafficking, membrane fusion events and gained particular interest for their role in host cell invasion by different pathogens, including *Plasmodium*.

The host laboratory performed mass spectrometry analysis of *P. falciparum* lipid rafts from trophozoites, early and late schizonts and sexual gametocytes, collecting parasite samples from synchronous cultures (figure 18). We used the four proteomic data sets to filter the predicted PPI network of *P. falciparum* and derive the stage-specific lipid-raft interactomes. In table IV we reported the numbers of proteins experimentally identified by mass spectrometry, the number of proteins and connections of the obtained interactomes.

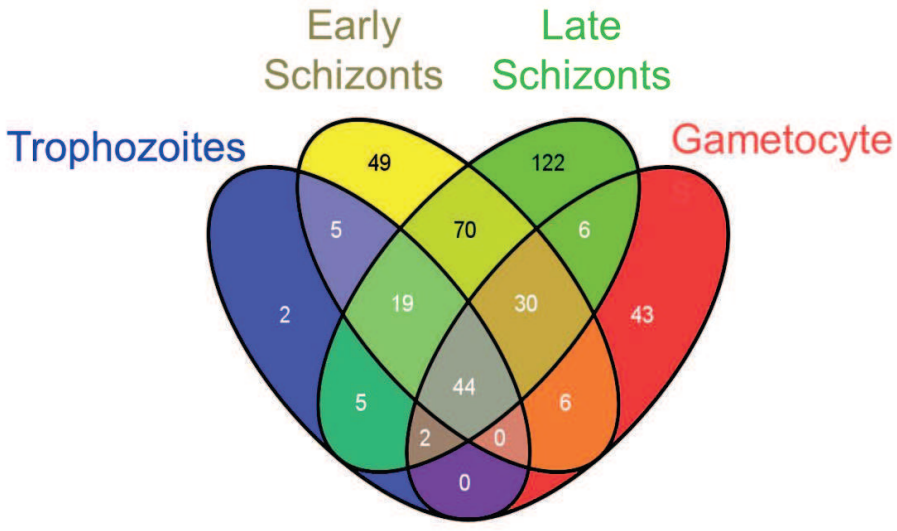


Figure 18. Summary and overlap of the mass spectrometry data of lipid rafts. Venn diagram of mass spectrometry data of lipid rafts from four *P. falciparum* developmental stages

Table IV. Numbers of the identifications and characteristics of the stage-related networks

	Experimentally identified proteins	*# proteins in the raft-related network*	*# connections*
Young trophozoites	77	70	214
Early schizonts	223	213	1159
Late schizonts	298	282	2141
Stage IV gametocytes	131	124	358

In all cases we found that the lipid raft-associated proteins are involved in highly connected networks. In order to assess the probability that this can be obtained by chance, we performed a random simulation test selecting ten protein sets, from the global PPI network, of the same size of proteomic data sets. Comparative analysis was performed calculating the Density values according to $D = \frac{2E}{N(N-1)}$ (where E is the number of connections observed in a given networks and $N(N-1)$ the number of all possible connections between N nodes), which provides a measure of the internal connectivity of a network. Results are shown in figure 19 as scatter dot plots. In all cases we found that the

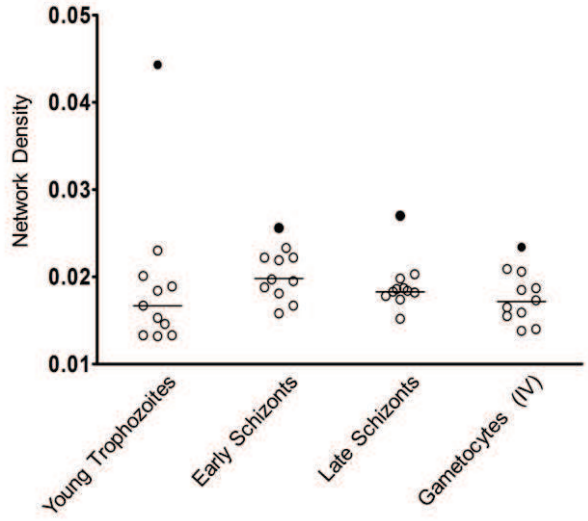


Figure 19. Simulation test. The network densities of lipid-raft PPI networks (empty dots) are compared with network densities of PPI networks obtained by randomly choosing ten protein data sets of equal size (empty dots).

mean density values (horizontal bars) of interactomes derived from randomly chosen proteins (empty dots) are lower than those obtained from the predicted interactomes (full dots).

4.6.a Analysis of dynamic interactions.

Once established that the stage-specific interactomes significantly differ from those obtained using random data sets, we proceeded with their characterization. One of our main interests, is to gain insights on the dynamics of membrane lipid rafts during the parasite development. To this aim, we applied a methodology (Srihari & Leong, 2012) that measure the percentage of conservation of PPIs.

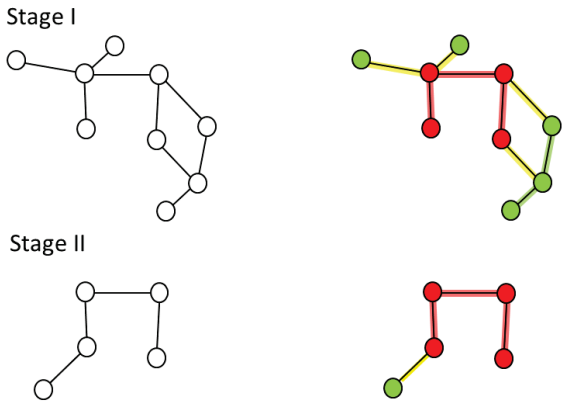


Figure 20. "Static" and "dynamic" classification. Stage-specific proteins (green nodes) are classified as "dynamic" (D), conserved proteins (red nodes) as "static" (S). Edges are classified as *D*, "static-dynamic" (SD) or S if they connect two D proteins (green edges), a S and a D protein (yellow edges) or two S proteins (red edges), respectively.

Based on comparative analysis of lipid raft proteomes of the four examined stages, a stage-specific protein is classified as "dynamic", while a protein conserved in at least two stages is considered as "static". Accordingly, PPIs are classified as "dynamic" (*D*) if they occur between two dynamic proteins, "static/dynamic" (*SD*) if they connect a "static" with a "dynamic" protein and "static" and (*S*) if the predicted interaction occurs between two "static" proteins (figure 20).

Stage	*# proteins*	*# interactions*	*Interaction type*		
			D	*SD*	*S*
Young Trophozoites	70	214	0	0	214 (100%)
Early schizonts	213	1159	61 (5.2%)	302 (26.1%)	796 (68.7%)
Late schizonts	282	2141	461 (21.5%)	909 (42.5%)	771 (36%)
Stage IV Gametocyte	124	358	23 (6.4%)	40 (11.2%)	295 (82.4%)

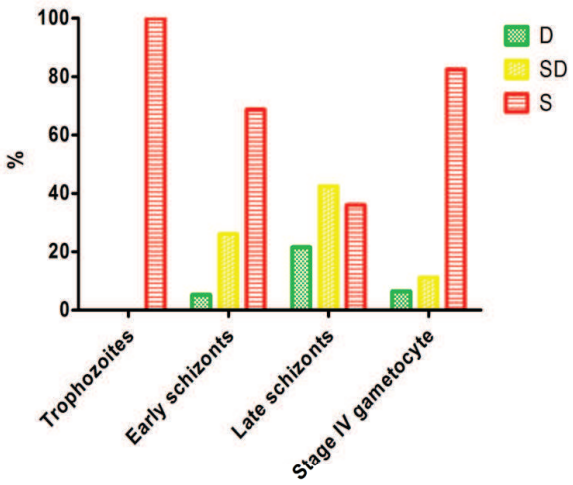


Figure 21. Dynamic of the *P. falciparum* lipid-raft interactomes. Results of PPI classification in the stage-specific interactomes are shown as histograms. For each interactome the percentage of D, SD and S PPIs is reported.

As shown in figure 21, the interactions classified as S are preponderant in the networks of early schizonts and gametocytes, whereas they represent 36% of total interactions in late schizonts. Interestingly, the trophozoite network contains only S interactions, indicating that no specific function is represented in lipid rafts of this stage. D interactions

are more represented in late schizonts (about 20% of total interactions), while they slightly exceed 5% in early schizonts and gametocytes. *SD* interactions imply that a "static" protein is connected with "dynamic" partners, suggesting that a conserved protein may be implicated in different processes depending on parasite developmental stage.

Taken together *SD* and *D* interactions, completely absent in the trophozoite network, represent about the 31% and 18% of the early schizonts and gametocyte networks respectively, and more than 60% in late schizonts. Thus, it appears that late schizont interactome, with a higher number of interactions established between stage-specific proteins, is more "dynamic" than the other stages considered. This can reflect the fact that late schizonts contain fully developed, highly specialized invasive forms (merozoites) characterized by specific organelles, necessary to establish contact and penetrate inside the host erythrocyte. In the following paragraphs, we integrated S, SD and D classification with functional and topological information.

4.6.b Cluster analysis.

In principle, proteins with functional relationships are highly connected and form distinct clusters in a PPI network.

Here we applied the AllegroMCODE software (Bader, et al, 2003) to identify clusters within each stage-specific *P. falciparum* network and assigned a functional annotation to the proteins populating each cluster. Examples of

clusters enriched in specific functional categories are presented in figure 22 where networks are represented as graphs and larger nodes indicate "dynamic" components.

The invasion-related proteins (green dots in figure 22) poorly populate the

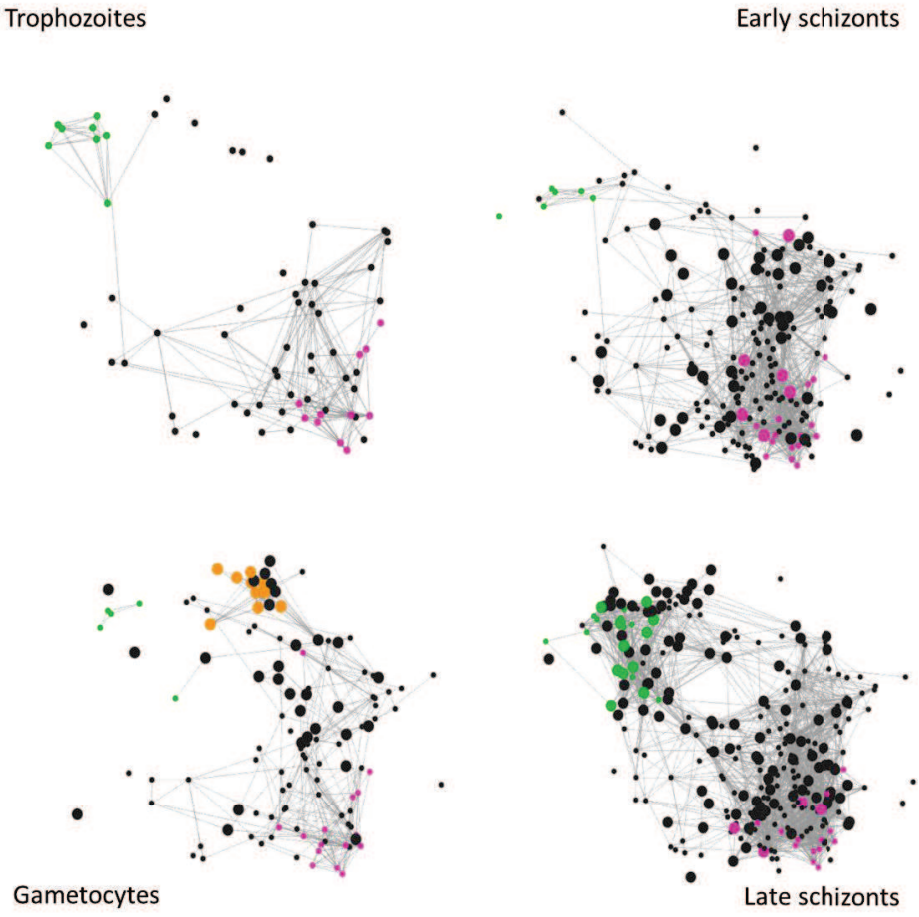


Figure 22. *P. falciparum* lipid-raft PPI networks. The stage-specific interactomes of lipid rafts are shown as graphs, where nodes represent proteins and edges their connections. Green nodes represent proteins annotated as "invasion related", while pink and orange nodes represent proteins involved in folding and egress/fertilization processes, respectively. Larger nodes correspond to proteins classified as "dynamic".

interactomes of trophozoites and early schizonts, while they form a large cluster in late schizonts. Further, proteins involved in early steps of gametogenesis (orange dots in figure 22), i.e. the egress and fertilization, form a distinct cluster in gametocyte network. Egress, is a process leading to the rupture of the erythrocyte membrane to enable gamete release and the successive fusion of female and male gametes (fertilization). These processes occur in the mosquito midgut, following an infected blood meal, although several involved proteins are already expressed at the gametocyte stage. In figure 22, we also highlighted proteins involved in folding (pink nodes), a category represented with a variable abundance in stage-specific interactomes, likely reflecting the weight that raft-mediated folding has in the different developmental stages. Interestingly, in early and late schizonts members of this protein family are recruited in lipid rafts in a stage-specific manner (dynamic nodes), while in trophozoites and gametocytes all detected chaperonins are implicated in conserved pathways (static nodes).

These examples show that cluster analysis, combined with functional annotation of protein components may clarify a specific biological context. This information can be also exploited to assign a role to proteins of unknown functions, residing in the same cluster. An example is shown in figure 23 where proteins implicated in either egress or fertilization are connected with four unknown “dynamic” proteins residing in the same cluster of gametocyte interactome (orange and black nodes respectively). These proteins may represent new candidates involved in the egress process and place them as suitable targets for further investigation.

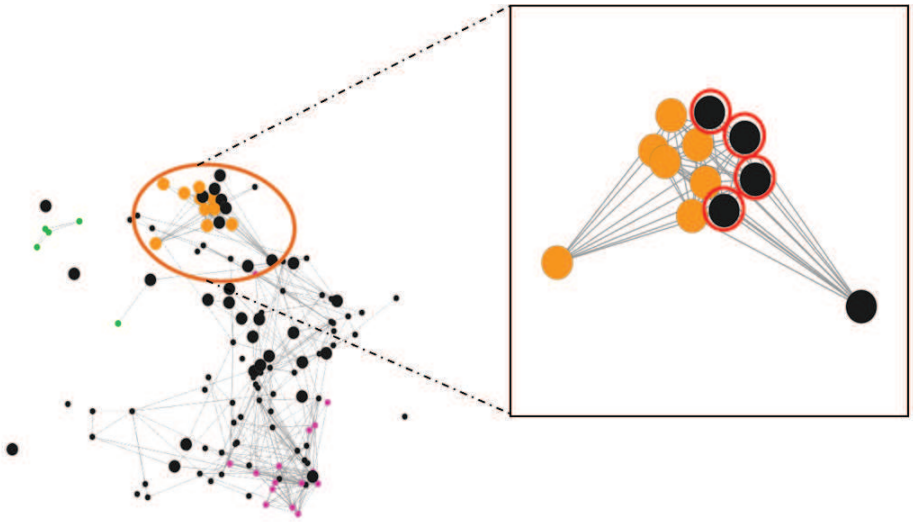


Figure 23. Egress/fertilization cluster in gametocyte interactome. A cluster of proteins involved in egress/fertilization (orange nodes) is highlighted in the gametocyte interactome (left panel). In the inset the four nodes rounded in red are proteins of unknown function.

4.6.c The lipid raft "functional core".

As shown above, lipid raft interactomes include proteins conserved between networks, suggesting the presence of a common pattern of interactions maintained across all the various stages. To define this "functional core", we decided to apply a network alignment to the four lipid raft networks (Wang & Gao, 2012; Bandyopadhyay, et al, 2006). We adopted the GASOLINE software (Micale, et al, 2014) that perform a multiple local alignment and evaluates its quality by measuring the average degree of conservation by an Index of Structural Conservation (ISC) ranging from 0 to 1.

We identified a high quality alignment with a conserved topology across the four stages (ISC=1), composed of 24 nodes (figure 24). When we examined this subnetwork by a functional point of view (see Table V), we observed that several interacting partners localize to either the parasitophorous vacuole

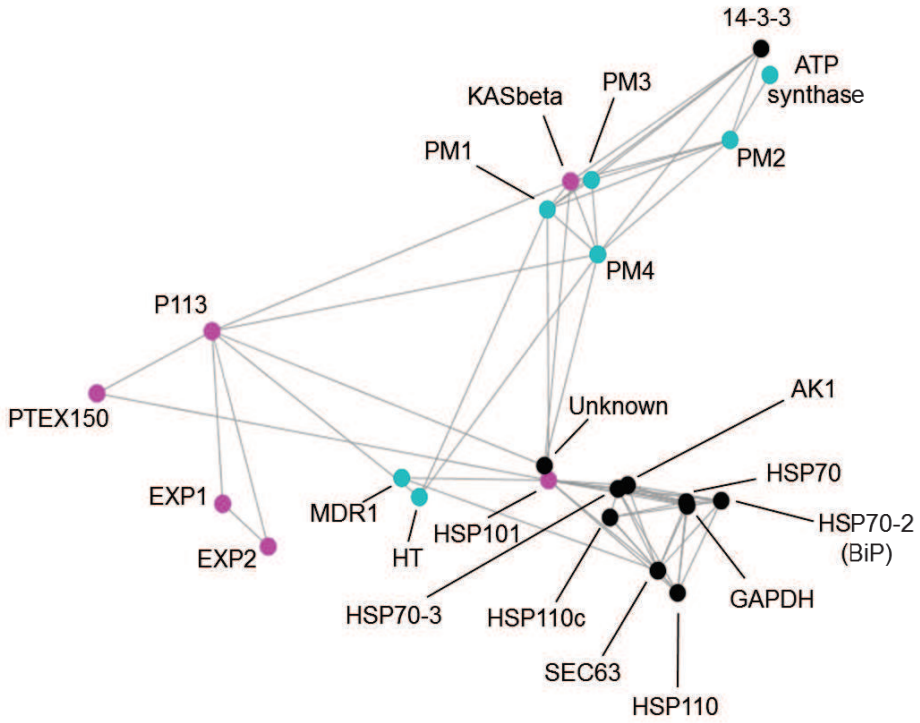


Figure 24. Lipid raft "functional core". A sub-network topologically conserved across the four interactomes was revealed by aligning the stage-specific PPI networks. Pink nodes represent proteins localized to the PVM, while light blue nodes represent FV membrane proteins.

membrane (PVM) (pink nodes in figure 24), or the food vacuole membrane (FVM) (light blue nodes in figure 24), subcellular compartments involved in processes important for parasite development inside the host erythrocyte.

The PVM that surrounds the parasite during the entire erythrocyte cycle is involved in protein sorting and trafficking to internal subcellular compartments or to the erythrocyte cytosol. Parasite proteins routed to the host cell compartment cross the PVM using a translocation machinery, also known as PTEX complex (de Koning-Ward, et al, 2009). Several subunits of the PTEX reside in the "functional core" subnetwork and are shown to interact between each other. The "functional core" subnetwork also includes numerous members of the heat shock protein family. These class of proteins have been demonstrated to be crucial under fever attacks in guaranteeing the correct protein folding (Pavithra, et al, 2004), but they are also enriched in the PVM (Nyalwidhe & Lingelbach, 2006) where they likely guide unfolding and refolding of proteins trafficked between membrane compartments.

Table V: protein of the "functional core" with possible evidence of their cellular localization.

Protein ID	Product description	Localization & Reference	
MAL13P1.540	heat shock protein 110, putative (HSP110)		
MAL8P1.69	14-3-3 protein (14-3-3I)		
PF07_0033	heat shock protein 110 (HSP110c)		
PF08_0054	heat shock protein 70 (HSP70)		
PF10_0086	adenylate kinase (AK1)		
PF11_0175	heat shock protein 101,chaperone protein ClpB2 (HSP101)	PVM	de Koning-Ward, et al, 2009; Riglar, et al, 2012
PF11_0224	circumsporozoite-related antigen,exported protein 1 (EXP1)	PVM	Riglar, et al, 2012

PF11_0351	heat shock protein 70 (HSP70-3)		
PF13_0102	translocation protein SEC63 (SEC63)		
PF14_0075	plasmepsin IV (PM4)	FV	Banerjee, et al, 2002
PF14_0076	plasmepsin I (PM1)	FV	Banerjee, et al, 2002
PF14_0077	plasmepsin II (PM1)	FV	Banerjee, et al, 2002
PF14_0078	plasmepsin III,histo-aspartic protease (PM3)	FV	Banerjee, et al, 2002
PF14_0201	surface protein P113 (P113)	PVM	Obando-Martinez, et al, 2009
PF14_0344	translocon component PTEX150 (PTEX150)	PVM	de Koning-Ward, et al, 2009
PF14_0598	glyceraldehyde-3-phosphate dehydrogenase (GAPDH)		
PF14_0615	ATP synthase (C/AC39) subunit, putative	FV*	Lamarque, et al, 2008
PF14_0678	exported protein 2 (EXP2)	PVM	Fischer, et al, 1998; Ito, et al, 2011
PFB0210c	hexose transporter (HT)	FV*	Slavic, et al, 2010
PFE1150w	multidrug resistance protein (MDR1)	FV	Cowman, et al, 1991
PFE1195w	karyopherin beta (KASbeta)	PVM*	Mohammed, et al, 2005
PFI0875w	heat shock protein 70 (HSP70-2), BiP		
PFI1270w	conserved *Plasmodium* protein, unknown function		

*predicted

Plasmepsins and the multi-drug resistance 1 (MDR1) (light blue nodes in figure 24) localize to the food vacuole membrane (FVM), where they are involved in the hemoglobin digestion. *Plasmodium* parasites engulfs part of the host cytoplasm through the invagination of the PVM and the generated vesicles release their content in the food vacuole (FV) (Lazarus, et al, 2008) where the hemoglobin digestion starts by plasmepsins (Banerjee, et al, 2002). The membrane pump MDR1, is then involved in translocating small peptides derived from the digestion of the hemoglobin that represent one of the major amino acid sources for the parasite. About 60-80% of the host hemoglobin is digested during *Plasmodium* infection (Francis, et al, 1997), suggesting a constant trafficking of vesicles from the PVM to the FV and thus a sort of constant interaction between these compartments, that may be mediated by membrane microdomains. Consistent with the functional interaction of these membrane compartments is the interconnection of their components observed in the "functional core" subnetwork.

The inspection of the "functional core" subnetwork suggests other possible candidates that may populate the functional interface between the PVM and FV. One of them is the Hexose Transporter (HT; PFB0210c), whose subcellular localization has not been determined in *P. falciparum* blood stages. Interestingly, the rodent *P. berghei* orthologs has been localized at the food vacuole (Slavic, et al, 2010). Overall these results underline the sensitivity of our interactome in describing the functional interactions between proteins.

5- CONCLUSIONS AND PERSPECTIVES

The reconstruction of the complex networks of functional and physical relationships between proteins is a central issue for the understanding of biological processes. For this reason, the prediction of interactomes is one of the most important challenges for both experimental and computational biologists. In the last ten years, a great amount of data has been produced by "high-throughput" approaches. The number of sequenced genomes, their quality in terms of completeness and the annotation of the gene repertoires significantly increased. In parallel, novel data sets have been acquired by transcriptomic approaches and mass spectrometry analyses on whole cell proteomes under specific conditions or at different developmental stages.

Computational methods to predict PPIs emerged in parallel with the first releases of completely sequenced genomes. The availability of this great amount of genomic data and the corresponding repertoires of predicted proteins made it possible to develop novel classes of predictive algorithms based on statistical methods.

In this work, we applied computational methods to obtain a new PPI network for the malaria parasite *P. falciparum*, with the aim to gain insights on the biological functions and PPIs taking place in membrane microdomains or lipid rafts. This subcellular compartment is involved in crucial cellular processes such as signal transduction, protein sorting and trafficking and membrane fusion events. Lipid raft characterization in *Plasmodium*-infected erythrocytes may provide new insights on parasite biology and potential candidates for drugs development.

Taking advantage of the available genomic, transcriptomic and proteomic, data, we applied a Bayesian approach to predict a probabilistic global interactome of *P. falciparum* and used this information to derive PPI networks of lipid raft-associated proteins in four different stages of the parasite development.

To this aim, we performed a re-assessment of two computational approaches usually used to predict PPIs: the phylogenetic profiling and the construction of the rosetta stone data.

Thanks to the increasing amount of available genomic data, phylogenetic profiling is one of the most promising computational approaches to predict PPIs. It is based on the detection of co-evolving proteins, which are often involved in the same biological pathways or in the same complex. This method relies on the comparison of proteins from a given genome with those from genomes in a reference set.

The two crucial points of this method are the selection procedure to include genomes in the reference set and the choice of the metric to compare the phylogenetic profiles. Redundancies in the reference set as well as the quality of the genome sequences affect the predictive performances of the method. A fortiori, the metric adopted to compare phylogenetic profiles clearly influences the prediction of true PPIs. Several studies demonstrated that, correlation based metrics provide the best results in detecting pairs of interacting proteins (Glazko, et al, 2004). These metrics, however, capture only linear relationships between variables, while mutual information reveals both linear and not linear associations. Recently, Szekely and collaborators proposed a novel statistics as an extension of the Pearson's correlation coefficient (Szekely, et al, 2007). They showed that distance correlation is equal to zero, if and only if the two

variables are independent, while it assumes positive values if any relationship exists between them.

We addressed the two issues regarding the reference set construction and the comparison of phylogenetic profiles, developing a new strategy of genome selection, based on the phylogenetic relationships between them, and applying the distance correlation as a novel measure of profile similarity. This results in an improvement of phylogenetic profiling in PPI prediction. Furthermore, we confirmed the influence of the reference set in terms of size and genome composition. Our strategy of genome selection allowed for the construction of an "outperforming" reference set, which was then used to derive the novel set of phylogenetic profiles for *P. falciparum*.

The same procedure was also adopted to construct the reference set (RS1) used to derive new data on rosetta stone proteins for *P. falciparum*. Although, the number of parasite proteins that we found to be encoded by fused genes in the reference set is comparable to those detected in previous studies, we obtained a strong improvement in the accuracy of the method.

Moreover, a further validation of the robustness of RS1, comes from an independent analysis of a potential "rosetta-stone protein" of the pathogen protozoan *G. duodenalis*. Database searches, using several portions of the protein *versus* RS1 suggested that a gene fusion event might be occurred in a bacterium followed by a gene-transfer event to the common ancestor of *G. duodenalis* and *E. hystolitica,* that shared the ecological niches of the human intestine.

As interacting proteins are very often found to be encoded by co-expressed genes, in parallel with the construction of novel genomic data sets, we exploited

available transcriptomic data to derive and compare new gene expression profiles covering both sexual and asexual stages.

Genomic and transcriptomic data were integrated and a global probability score was obtained for each predicted PPI. We showed that the predictive performance of this global score, strongly outperforms that obtained by the single scores. Then, we extracted a "high-accuracy" set of PPIs with a global score corresponding to a ratio of TP/FP higher than 10 and containing more than the 88% percent of *P. falciparum* proteins. The resulting PPI network includes all possible interactions irrespective of protein co-occurrences in terms of timing of expression or cellular localization. To overcome this problem, we filtered the PPI network with proteomic data on lipid rafts-associated proteins in four developmental stages of *P. falciparum*, obtained by the host lab. The functional and topological characteristics of the four stage-specific interactomes were explored using different methods. We observed that lipid raft PPI interactome undergoes to a continuous remodeling during the parasite development. We also observed that proteins involved in stage-specific processes, like invasion and egress, form distinct clusters. The fact that these clusters include proteins whose function is still unknown, can be exploited to predict their possible biological roles.

Furthermore, the alignment of the four PPI networks revealed a "functional core" of raft-associated proteins forming a subnetwork topologically conserved in the four PPI networks. This subnetwork is formed by proteins involved in two fundamental processes of the parasite cell: hemoglobin digestion and protein sorting/trafficking.

This work represents an improvement in the quality and accuracy of PPI network prediction for *P. falciparum*, and opens new perspectives in the investigation of specific biological questions.

The possibility to filter the network using diverse proteomic data enables to include spatial and temporal information and gives the opportunity to analyze the dynamics of subcellular compartments during developmental phases or under different *stimuli*. This may suggest new drug/vaccine candidates to be submitted to experimental validation.

Furthermore, the genomic data sets that we produced may have other interesting applications. The comparative analysis of parasite and host proteins by phylogenetic profiling may provide insights on parasite-host interactions that play important role in several crucial processes, like sequestration of infected erythrocyte, host cell remodeling by the generation of novel membrane compartments, invasion and egress from the host cell (preliminary analyses are presented in Part III).

Furthermore, novel phylogenetic profiles and rosetta stone data may have a more general application to gain insights on gene evolution (an example is presented in Part III).

6 – REFERENCES

- Abkarian M, Massiera G, Berry L, Roques M, Braun-Breton C. A novel mechanism for egress of malaria parasites from red blood cells. *Blood*, 2011; 117(15): 4118-4124.
- Allen JA, Halverson-Tamboli RA, Rasenick MM. Lipid raft microdomains and neurotransmitter signaling. *Nature Reviews Neuroscience*, 2007; 8(2): 128-140.
- Bader GD & Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 2003; 4: 2.
- Balu B. Moving "forward" in *Plasmodium* genetics through a transposon-based approach. *Journal of Tropical Medicine*, 2012; 828210.
- Bandyopadhyay S, Sharan R, Ideker T. Systematic identification of functional orthologs based on protein network comparison. *Genome Research*, 2006; 16(3): 428-435.
- Banerjee R, Liu J, Beatty W, Pelosof L, Klemba M, Goldberg DE. Four plasmepsins are active in the *Plasmodium falciparum* food vacuole, including a protease with an active-site histidine. *Proceedings of the National Academy of Sciences of the United States of America*, 2002; 99(2): 990-995.
- Barabasi AL, Albert R. Emergence of scaling in random networks. *Science*, 1999; 286(5439): 509-512.
- Becker S & Le Cun Y. Improving the convergence of back-propagation learning with second-order methods. *Proceedings of the 1988 Connectionist Models Summer School, 1988*; pages 29-37.
- Bentley DR, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 2008; 456(7218): 53-59.
- Blackman MJ. Malarial proteases and host cell egress: an "emerging" cascade. *Cellular Microbiology*, 2008; 10(10): 1925-1934.

- Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biology*, 2003; 1(1): E5.
- Chaves LF, Koenraadt CJ. Climate change and highland malaria: fresh air for a hot debate. *The Quarterly Review of Biology*, 2010; 85(1): 27-55.
- Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. *Science*, 2006; 311(5765): 1283-1287.
- Cilingir G, Broschat SL, Lau AOT. ApicoAP: the first computational method for identifying apicoplast targeted proteins in multiple species of apicomplexan. *PLoS ONE*, 2012; 7(5): e36598.
- Cowman AF, Karcz S, Galatis D, Culvenor JG. A P-glycoprotein homologue of *Plasmodium falciparum* is localized on the digestive vacuole. *The Journal of Cell Biology*, 1991; 113(5): 1033-42.
- Cowman AF & Crabb BS. Invasion of red blood cells by malaria parasites. *Cell*, 2006; 124(4): 755-766.
- Date SV, Marcotte EM. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnologies*, 2003; 21(9):1055-1062.
- Date SV & Stoeckert CJ. Computational modeling of the *Plasmodium falciparum* interactome reveals protein function on a genome-wide scale. *Genome Research*, 2006; 16(4): 542-549.
- Deane CM, Salwiński L, Xenarios I, Eisenberg D. Two methods for assessment of the reliability of high throughput Observations. *Molecular & Cellular Proteomics*, 2002; 1: 349-356.
- De Koning-Ward TF, Gilson PR, Boddey JA, Rug M, Smith BJ, Papenfuss AT, Sanders PR, Lundie RJ, Maier AG, Cowman AF, Crabb BS. A newly discovered protein export machine in malaria parasites. *Nature*, 2009; 459(7249): 945-959.
- Di Girolamo F, Raggi C, birago C, Pizzi E, Lalle M, Picci L, Pace T, Bachi A, de Jong J, Janse CJ, Waters AP, Sargiacomo M, Ponzi M. *Plasmodium* lipid rafts contain proteins implicated in vesicular

trafficking and signalling as well as members of the PIR superfamily, potentially implicated in host immune system interactions. *Proteomics*, 2008; 8(12): 2500-2513.

- Ding B, Gentleman R, Carey V. bioDist: different distance measures. R package source http://bioconductor.org/biocLite.R bioclite ("bioDist"), 2010.
- Enault F, Suhre K, Abergel C, Poirot O, Claverie JM. Annotation of bacterial genomes using improved phylogenomic profiles. *Bioinformatics*, 2003; 19 Suppl 1: 105-107.
- Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 1999; 402(6756): 68-90.
- Fischer K, Marti T, Rick B, Johnson D, Benting J, Baumeister S, Helmbrecht C, Lanzer M, Linglebach K. Characterization and cloning of the gene encoding the vacuolar membrane protein EXP-2 from *Plasmodium falciparum. Molecular and Biochemical Parasitology,* 1998; 92(1): 47-57.
- Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, Witney AA, Wolters D, Wu Y, Gardner MJ, Holder AA, Sinden RE, Yates JR, Carucci DJ. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature*, 2002; 419(6906): 520-526.
- Foth BJ, Ralph SA, Tonkin CJ, Struck NS, Fraunholz M, Roos DS, Cowman AF, McFadden GI. Dissecting apicoplast targeting in the malaria parasite *Plasmodium falciparum*. *Science*, 2003; 299(5607): 705-708.
- Franceschini A, Szlarczyk D, Franklid S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 2013; 41(Database issue): D808-D815.
- Francis SE, Sullivan DJ, Goldberg DE. Hemoglobin metabolism in the malaria parasite *Plasmodium falciparum*. *Annual Review of Microbiology*, 1997; 51:97-123.

- Gardner MJ, Hall N, Fung E, White O, Barriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Keys S, Chan M, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DMA, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 2002; 419(6906): 498-511.
- Gibson G. Microarray analysis: genome-scale hypothesis scanning. *PLoS Biology*, 2003; 1(1): E15.
- Glazko GV& Mushegian AR. Detection of evolutionary stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biology*, 2004; 5(5): R32.
- Gubler DJ, Reiter P, Ebi KL, Yep W, Nasci R, Patz JA. Climate variability and change in the United States: potential impacts on vector- and rodent-borne diseases. *Environmental Health Perspectives*, 2001; 109 (Suppl. 2): 223-233.
- Guiguemde WA, Shelat AA, Garcia-Bustos JF, Diagana TT, Gamo FJ, Guy RK. Global phenotypic screening for anitmalarials. *Chemistry & Biology*, 2012; 19(1): 116-129.
- Halliez MCM & Buret AG. Extra-intestinal and long term consequences of *Giardia duodenalis* infections. *World Journal of Gastroenterology*, 2013; 19(47): 8974-8985.
- Hu G, Cabrera A, kono M, Mok S, Chaal BK, Haase S, Engelberg K, Cheemadan S, Spielmann T, Preiser PR, Gilberger TW, Bozdech Z. Transcriptional profiling of growth perturbations of the human malaria parasite *Plasmodium falciparum*. *Nature Biotechnologies*, 2010; 28(1): 91.98.
- Inselburg J. Stage-specific inhibitory effect of cyclic AMP on asexual maturation and gametocyte formation of *Plasmodium falciparum*. *Journal of Parasitology*, 1983; 69(3): 592-597.

- Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 2001; 98(8): 4569-4575. doi: 10.1073/pnas.061034498
- Jansen R, Greenbaum D, Gerstein M. Relating whole-genome expression data with protein-protein interactions. *Genome Research*, 2002; 12(1): 37-46.
- Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. A Bayesian network approach for predicting protein-protein interactions from genomic data. *Science*, 2003; 302(5644): 449-453.
- Joe H. Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association*, 1989; 84(405): 157-164.
- Kanehisa M & Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 2000; 28(1): 27-30.
- Koshino I & Takakuwa Y. Disruption of lipid rafts by lidocaine inhibits erythrocyte invasion by *Plasmodium falciparum*. *Experimental Parasitology*, 2009; 123(4): 381-383.
- LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C, Fields S, Hughes RE. A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature*, 2005; 438(7064): 103-107.
- Lamarque M, Tastet C, Poncet J, Demettre E, Jouin P, Vial H, Dubremetz JF. Food vacuole proteome of the malaria parasite *Plasmodium falciparum*. *Proteomics. Clinical Applications*, 2008; 2(9): 1361-1374.
- Lasonder E, Ishihama Y, Andersen JS, Vermunt AMW, Pain A, Sauerwein RW, Eling WMC, Hall N, Waters AP, Stunnenberg HG, Mann M. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature*, 2002; 419(6906): 537-542.
- Lazarus MD, Schneider TG, Taraschi TF. A new model for hemoglobin ingestion and transport by the human malaria parasite *Plasmodium falciparum*. *Journal of Cell Science*, 2008; 121 (Pt11): 1937-1949.

- Lee I, Date SV, Adai AT, Marcotte EM. A probabilistic functional network of yeast genes. *Science*, 2004; 306(5701): 1555-1558.
- Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch K, Haynes D, De la Vega P, Holder AA, Batalov S, Carucci DJ, Winzeler EA. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, 2003; 301(5639): 1503-1509.
- Lim L & McFadden GI. The evolution, metabolism and functions of the apicoplast. *Philosophical Transaction of the Royal society of London. Series B, Biological Sciences*, 2010; 365(1541): 749-763.
- Lingwood D & Simons K. Lipid Rafts as a membrane-organizing principle. *Science*, 2010; 327(5961): 46-50.
- Llinas M, Bozdech Z, Wong ED, Adai AT, DeRisi JL. Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. *Nucleic Acids Research*, 2006; 34(4): 1166-1173.
- Marcotte CJV, Marcotte EM. Predicting functional linkages from gene fusions with confidence. *Applied Bioinformatics*, 2002; 1(2): 93-100.
- Martin W & Herrmann RG. Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiology*, 1998; 118 (1): 9-17.
- McFadden GI, Reith ME, Munholland J, Lang-Unnasch N. Plastid in human parasites. *Nature*, 1996; 381(6582): 482.
- Micale G, Continella A, Ferro A, Giugno R, Pulvirenti A. GASOLINE: a Cytoscape app for multiple local alignment of PPI networks. *F1000Research*, 2014; 3:140.
- Mohammed A, Kishore S, Patra KP, Dasaradhi PV, Malhotra P, Chauhan VS. Identification of karyopherin beta as an immunogenic antigen of the malaria parasite using immune mice and human sera. *Parasite Immunology*, 2005; 27(5): 197-203.
- Muller J, & Hemphill A. New approaches for the identification of drug targets in protozoan parasites. *International Review of Cell and Molecular Biology*, 2013; 301: 359-401.

- Murphy SC, Hiller NL, Harrison T, Lomasney JW, Mohandas N, Haldar K. Lipid rafts and malaria parasite infection of erythrocytes. *Molecular Membrane Biology*, 2006; 23(1): 81-88.
- Nagalakshmi U, Waern K, Snyder M. RNA-Seq: a method for comprehensive transcriptome analysis. *Current Protocols in Molecular Biology*, 2010; 4(4): 1-13.
- Newman MEJ. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2006; 103(23): 8577-8582.
- Nyalwidhe J & Lingelbach K. Proteases and chaperones are the most abundant proteins in the parasitophorous vacuole of *Plasmodium falciparum*-infected erythrocyte. *Proteomics*, 2006; 6(5): 1563-1573.
- Obando-Martinez AZ, Curtidor H, Arevalo-Pinzon G, Vanegas M, Vizcaino C, Patarroyo MA, Patarroyo ME. Conserved high activity binding peptides are involved in adhesion of two detergent-resistant membrane-associated merozoite proteins to red blood cells during invasion. *Journal of Medical Chemistry*, 2010; 53(10): 3907-3918.
- Pavithra SR, Banumathy G, Joy O, Tatu U. Recurrent fever promotes *Plasmodium falciparum* development in human erythrocytes. *Journal of Biological Chemistry*, 2004; 279(45): 46692-46699.
- Odolini S, Gautret P, Parola P. Epidemiology of imported malaria in the Mediterranean region. *Mediterranean Journal of Hematology and Infectious diseases*, 2012; 4(1): e2012031.
- Otto TD, Wilinski D, Assefa S, Keane TM, Sarry LR, Bohme U, Lemieux J, Barrell B, Pain A, Barriman M, Newbold C, Llinas M. New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. *Molecular Microbiology*, 2010; 76(1): 12-24.
- Pearson WR & Lipman DJ. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Science of the United States of America*, 1988; 85(8): 2444-2448.
- Pei Y, Tarun AS, Vaughan AM, Herman RW, Sollman JMB, Erickson-Wayman A, Kappe SH. *Plasmodium* pyruvate dehydrogenase activity is only essential for the parasite's progression from liver infection to blood infection. *Molecular Microbiology*, 2010; 75(4): 957-971.

- Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America, 1999;* 96(8): 4285-4288.
- Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, Jensen LJ, von Mering C, Bork P. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Research*, 2012; 40(Database issue): D284-D289.
- R Development Core Team (2012). R: a language and environment for statistical computing. R foundation for statistical computing, Vienna Austria.
- Rajendran L, Knolker HJ, Simons K. Subcellular targeting strategies for drug design and delivery. *Nature reviews. Drug Discovery*, 2010. 9(1): 29-42.
- Ramaprasad A, Pain A, Ravas T. Defining the protein interaction network of human malaria parasite *Plasmodium falciparum*. *Genomics*, 2012; 99(2): 69-75.
- Riglar DT, Rogers KL, Hanssen E, Turnbull L, Bullen HE, Charnaud SC, Przyborski J, Gilson PR, Whitchurch CB, Crabb BS, Baum J. Spatial association with PTEX complexes defines regions for effector export into *Plasmodium falciparum*-infected erythrocytes. *Nature communications*, 2013; 4:1415.
- Ryan U & Cacciò SM. Zoonotic potential of Giardia. *International Journal of Parasitology*, 2013; 43(12-13): 943-956.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 2003; 13(11): 2498-2504.
- Slavic K, Straschil U, Reininger L, Doering C, Morin C, Tewari R, Krishna S. Life cycle studies of the hexose transporter of *Plasmodium* species and genetic validation of their essentiality. *Molecular Micorbiology*, 2010; 75(6): 1402-1413.

- Smalley Mr & Sinden RE. *Plasmodium falciparum* gametocytes: their longevity and infectivity. *Parasitology*, 1977; 74(1): 1-8.
- Smith TF & Waterman MS. Identification of common molecular subsequences. *Journal of Molecular Biology*, 1981; 147(1): 195-197.
- Srihari S & Leong HW. Temporal dynamics of protein complexes in PPI networks: a case study using yeast cell cycle dynamics. *BMC Bioinformatics*, 2012; 13 Suppl 17: S16.
- Su X, Ferdig MT, Huang Y, Hyynh CQ, Liu A, You J, Wootton JC, Wellems TE. A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science*, 1999; 286(5443): 1351-1353.
- Szekely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 2007; 35(5): 2769-2794.
- Uetz P, Glot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinlvasan M, Pochart P, Quareshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 2000; 403(6770): 623-627.
- Vieira FS, Correa G, Einicker-Lamas M, Coutinho-Silva R. Host-cell lipid rafts: a safe door for micro-organisms? *Biology of the Cell*, 2010; 102(7): 391-407.
- Von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 2002; 417(6887): 399-403.
- Wang B & Gao L. Seed selection strategy in global network alignment without destroyin the entire structures of functional modules. *Proteome Science*, 2012; 10 Suppl. 1: S16.
- Watkins RR, & Eckmann L. Treatment of giardiasis: current status and future directions. *Current Infectious Disease Reports*, 2014; 16(2): 396.
- Watts DJ & Strogatz SH. Collective dynamics of "small-world" networks. *Nature*, 1998; 393(6684): 440-442.

- Wilson RJM, Denny PW, Preiser Pr, Rangachari K, Roberts K, Roy A, Whyte A, Strath M, Moore DJ, Moore PW, Williamson DH. Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*. Journal of molecular Biology, 1996; 261 (2): 155-172.
- WHO, World Malaria Report 2014.
- Young JA, Fivelman QL, Blair PL, de la Vega P, Le Roch KG, Zhou Y, Carucci DJ, Baker DA, Winzeler EA. The *Plasmodium falciparum* sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Molecular and Biochemical Parasitology*, 2005; 143(1): 67-79.
- Zahiri J, Bozorgmehr JH, Masoudi-Nejad A. Computational prediction of protein–protein interaction networks: algorithms and resources. *Current Genomics*. 2013;14(6):397-414.
- Zhang X, Zhao J, Hao JK, Zhao XM, Chen L. Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. *Nucleic Acids research*, 2015; 43(5): e31.
- Zuccala ES & Baum J. Cytoskeletal and membrane remodeling during malaria parasite invasion of the human erythrocyte. *British Journal of Haematology*, 2011; 154(6):680-689.
- Zuegge J, Ralph S, Schmuker M, McFadden GI, Schneider G. Deciphering apicoplast targeting signals – feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. *Gene*, 2001; 280(1-2): 19-26.

Acknowledgements

I would like to start thanking my PIs Dr. Elisabetta Pizzi and Dr. Marta Ponzi who decided to invest time, energies and resources on me and Dr. Erica Pasini that joined them in this role. Mainly I thank them to be my mentors.
I thank Evimalar (European Virtual Institute for Malaria Research) that, in agreement with the European Molecular Biology Laboratory (EMBL) and Università degli Studi di Milano, gave me the possibility to be part of its PhD Programme. It was an experience of science and life.
I thank my university advisor Dr. David Horner for the support.
I naturally thank Ponzi's lab people from which I was welcomed like a family. A special thinking for Tommaso Anna, and Federica.

During last four years my life was deeply enriched.
But what it will come is the best.

PART II

The FAD-dependent glycerol-3-phosphate dehydrogenase of Giardia duodenalis: an unconventional enzyme that interacts with the g14-3-3 and it is a target of the antitumoral compound NDBHEX.

Marco Lalle, Serena Camerini, Serena Cecchetti, Renata Finelli, Gabriella Sferra, Joachim Müller, Giorgio Ricci and Edoardo Pozio

Journal Name:	Frontiers in Microbiology
ISSN:	1664-302X
Article type:	Original Research Article
First received on:	04 Mar 2015
Revised on:	26 Apr 2015
Frontiers website link:	www.frontiersin.org

The FAD-dependent glycerol-3-phosphate dehydrogenase of *Giardia duodenalis:* an unconventional enzyme that interacts with the g14-3-3 and it is a target of the antitumoral compound NDBHEX

Marco Lalle[1*], Serena Camerini[2], Serena Cecchetti[2], Renata Finelli[1], Gabriella Sferra[1], Joachim Müller[3], Giorgio Ricci[4] and Edoardo Pozio[1].

[1]Department of Infectious, Parasitic and Immunomediated Diseases and [2]Department of Cell Biology and Neurosciences, Istituto Superiore di Sanità, viale Regina Elena 299, 00161 Rome, Italy; [3]Institute of Parasitology, Vetsuisse Faculty, University of Berne, Länggass-Strasse 122, CH-3012 Berne, Switzerland; [4]Department of Sciences and Chemical Technologies, University of Rome "Tor Vergata", via della Ricerca Scientifica 1, 00133 Rome, Italy.

Correspondence:

Dr. Marco Lalle
Istituto Superiore di Sanità
Department of Infectious, Parasitic and Immunomediated Diseases,
Gastrointestinal and Tissue Parasitic Diseases,
Viale Regina Elena 299,
00161 Rome, Italy
marco.lalle@iss.it

Abstract

The flagellated protozoan *Giardia duodenalis* is a worldwide parasite causing giardiasis, an acute and chronic diarrheal disease. Metabolism in *G. duodenalis* has a limited complexity thus making metabolic enzymes ideal targets for drug development. However, only few metabolic pathways (i.e, carbohydrates) have been described so far. Recently, the parasite homolog of the mitochondrial-like glycerol-3-phosphate dehydrogenase (gG3PD) has been identified among the interactors of the g14-3-3 protein. G3PD is involved in glycolysis, electron transport, glycerophospholipids metabolism and hyperosmotic stress response, and is emerging as promising target in tumor treatment. In this work, we

demonstrate that gG3PD is a functional flavoenzyme able to convert glycerol-3-phosphate into dihydroxyacetone phosphate and that its activity and the intracellular glycerol level increase during encystation. Taking advantage of co-immunoprecipitation assays and deletion mutants, we provide evidence that gG3PD and g14-3-3 interact at the trophozoite stage, the intracellular localization of gG3PD is stage dependent and it partially co-localizes with mitosomes during cyst development. Finally, we demonstrate that the gG3PD activity is affected by the antitumoral compound 6-(7-nitro-2,1,3-benzoxadiazol-4-ylthio)hexanol (NBDHEX), that results more effective *in vitro* at killing *G. duodenalis* trophozoites than the reference drug metronidazole. Overall, our results highlight the involvement of gG3PD in processes crucial for the parasite survival thus proposing this enzyme as target for novel anti-giardial interventions.

Running title: The gG3PD of *Giardia duodenalis.*

Keywords: *Giardia duodenalis*, FAD-dependent glycerol-3-phoshate dehydrogenase, 14-3-3 protein, energy metabolism, mitosome, encystation, NBDHEX, nitroreduction.

48 **Introduction**

49 The flagellated protozoan *Giardia duodenalis* (syn. *lamblia*, *intestinalis*) is a parasite of the upper part
50 of small intestine of mammals, including humans. Infection with *G. duodenalis* causes giardiasis, one
51 of the most common foodborne and waterborne gastroenteric diseases (Ryan and Cacciò, 2013; Halliez
52 and Buret, 2013). The parasite has a simple two-stages life cycle consisting of the trophozoite, that
53 replicates and colonizes the host intestine causing symptoms, and the cyst, the environmentally
54 resistant stage that is spread with feces and is responsible for transmission of the infection. Generally,
55 infection is acquired by ingestion of cysts in contaminated water and food or by the fecal-oral route
56 (Ryan and Cacciò, 2013). Clinical symptoms of giardiasis can vary from asymptomatic infection to
57 acute and chronic diarrhea, ultimately leading to chronic post-infectious gastrointestinal complications,
58 including irritable bowel syndrome and chronic fatigue (Halliez and Buret, 2013). Up to date, no
59 human vaccine for giardiasis is available and treatment relies only on a limited panel of effective
60 approved drugs. Nitroheterocyclics, such as metronidazole (MTZ) and nitazoxanide (NTZ), are the
61 antigiardial drugs of choice. Unfortunately, treatment failure has been reported in 10-20% of cases and
62 strains resistant to different compounds have been either clinically isolated or induced *in vitro* (Lalle,
63 2010; Watkins and Eckmann, 2014). In this scenario, alternative, safe and effective therapies are
64 required.

65 *G. duodenalis* has a peculiar energy metabolism. It is a microaerophilic organism that, instead of
66 mithocondria, contains mitosomes, highly reduced mitochondria-derived organelles which sole
67 function seems to be restricted to Fe-S cluster biosynthesis (Tovar et al., 2003; Jedelsky et al., 2011).
68 Energy is then generated by substrate level phosphorylation and fermentation occurring in the
69 cytoplasm or at the inner side of the plasma membrane (Adam, 2001). In terms of sequence similarity,
70 these metabolic pathways of *G. duodenalis* consist of a mixture of eukaryote-like and bacteria-like
71 enzymes. Therefore, energy and intermediate metabolism of *G. duodenalis* have been shown to provide
72 opportunities to identify novel and effective compounds, as well as potentially interesting targets by
73 means of high-throughput drug screening and target-based drug design (Müller and Hemphill, 2013;
74 Watkins and Eckmann, 2014).

75 We recently detected a putative glycerol-3-phosphate dehydrogenase/flavin-dependent oxidoreductase
76 (gG3PD, GL50803_16125) of *G. duodenalis* among proteins interacting with the parasite 14-3-3
77 isoform, g14-3-3 (Lalle et al., 2012). The 14-3-3s are a family of highly conserved eukaryotic
78 phoshoserine/phosphoteronine-binding proteins which participate to the regulation of key cellular
79 processes by direct interaction with hundreds of target proteins (Gardino and Yaffe, 2011; Kleppe et

al., 2011). The characterized g14-3-3 interactome provides evidences that g14-3-3 can be involved in parasite energy metabolism, as supported by the identification and confirmation of components of both the glycolytic/gluconeogenetic pathway and pyruvate metabolism (Lalle et al., 2012).

Flavin adenine dinucleotide (FAD)-dependent glycerol-3-phosphate dehydrogenase (G3PD, EC 1.1.5.3) is a key enzyme at the crossroad of glycolysis, redox and fatty acid metabolism, in both prokaryotes and eukaryotes. G3PD catalyzes the oxidation of glycerol-3-phosphate (G3P) to dihydroxyaceton phosphate (DHAP) with simultaneous reduction of FAD to $FADH_2$ and transfer of electrons to quinones (e.g. ubiquinone) (Unden and Bongaerts, 1997; Mráček et al., 2013). In eukaryotes, a single subunit enzyme (mG3PD, 69-75kDa) is strongly associated, as peripheral protein, with the outer face of the inner mitochondrial membrane (Janssen et al., 2002; Mráček et al., 2013). The mG3PD has multiple functions: i) it forms the glycerophosphate shuttle in combination with the cytosolic NADH-dependent G3PD (cG3PD, EC 1.1.1.8) to re-oxidize the cytosolic NADH produced by glycolysis; ii) it is part of the mitochondrial respiratory electron transport chain (ETC) channeling electron to quinone pool and bypassing Complex I; iii) it regulates the cytosolic level of G3P (Bell and Coleman, 1980; Mráček et al., 2013). Prokaryotes harbor two distinct membrane-associated FAD-dependent G3PDs, both necessary for bacterial growth in presence of glycerol or G3P as sole carbon source, and represent key primary dehydrogenases transferring reducing equivalents to a short respiratory ETC with different terminal reductases and electron acceptor (O_2, nitrate or fumarate) (Unden and Bongaerts, 1997). The homodimeric GlpD is associated with the cytoplasmic side of plasma membrane, shows 30-33% homology with eukaryotic mG3PD, and is expressed under aerobic conditions when O_2 is the terminal acceptor (Waltz et al., 2002; Yeh et al., 2008). The heterotrimeric GlpACB is induced under anaerobic conditions, with fumarate as terminal acceptor (Cole et al., 1988; Varga and Weiner, 1995): it forms a functional-associated complex with fumarate reductase and contributes to generate a proton gradient across the membrane via an associated ETC (Miki and Wilson, 1978). The GlpAC heterodimer (62 and 41 kDa, respectively) is the soluble catalytic subunit, containing the FAD and FMN binding sites (Cole et al., 1988). The GlpB (44 kDa) subunit, that contains a ferrodoxin-type (4Fe-4S) cluster binding motif, anchors the GlpAC dimer to the inner cytoplasmic membrane and mediate the transfer of reducing equivalent to the menaquinone pool and finally to the fumarate reductase (Varga and Weiner, 1995).

The G3PDs are also involved in pathogenicity as a source of reactive oxygene species (ROS). The mG3PD has been implied in the establishment of a pro-oxidative environment that promotes the fast growth of undifferentiated tumors (Chowdhury et al., 2007; Mráček et al., 2013), whereas the oxidase

activity of GlpD of *Mycoplasma* spp. seems to be crucial for the pathogenicity, leading to high levels of H_2O_2 and host cell damage (Großhennig et al., 2013).

In the present work, we investigated gG3PD expression, activity, and cellular localization in the *G. duodenalis* trophozoite stage and during the encystation process. Moreover, we demonstrated that the antitumoral compound (NBDHEX) is an effective antigiardial compound that negatively affected gG3PD activity. Overall, our results pointed out the role of gG3PD in biological processes crucial for the parasite survival, thus suggesting that this enzyme could represent a good candidate for targeted antigiardial interventions.

Materials and Methods

Chemicals. The 6-(7-nitro-2,1,3-benzoxadiazol-4-ylthio)hexanol (NBDHEX) was synthetized as previously described (Ricci et al., 2005). Metronidazole (MTZ), flavin adenine dinucleotide disodium salt (FAD), sn-glycerol-3-phosphate (g3p), phenazine methosulfate (PMS), 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) were from Sigma-Aldrich (St. Louis, Missouri, USA).

Parasite cultivation, differentiation and transfection. The *G. duodenalis* isolate WB clone C6 (WB-C6) was axenically grown in TYI-S-33 medium at 37 °C and differentiation into cyst (encystation) was induced in TYI-S-33 medium containing 5 mg/ml of bovine bile at pH 7.8 for the indicated time (Schupp et al., 1988). Parasites were collected by chilling tubes on ice and centrifugation at 800 × *g*. Transgenic lines were generated by trophozoite electroporation with 15 μg of plasmid DNA and selected and maintained under constant selection with 100 μM puromycin (Invivogen, Toulouse, France).

Vector construction. *Escherichia coli* BL21-DE3 competent cells were used. The full-length sequence of gG3PD (accession number GL50803_16125), the gG3PD N-terminal (nucleotide 3-1590; gG3PD_N) and the C-terminal half (nucleotide 1569-3333; gG3PD_C), were PCR amplified from the *G. duodenalis* WB-C6 genomic DNA, prepared as previously described (Lalle et al., 2012). The primers used and their combinations are reported in Table 1. Reactions were performed on a T-Personal Thermocycler (Biometra, Göttingen, Germany) using 100 ng of gDNA, 10 units of high fidelity Pfu turbo DNA polymerase (Agilent Technologies, Santa Clara, CA, USA), 50 μM dNTP, 20 pmols of each primer in 50 μl of reaction mixture. Amplification conditions were: 1 cycle at 95 °C for 2 min; 30

cycles at 95 °C for 30 sec, 58 °C for 30 sec and 72 °C for 30 sec; and 1 cycle at 72 °C for 7 min. PCR fragments were cloned either in the *BamHI/PspOMI* digested pTUB-FLAG_HApac vector (Lalle et al., 2012) for expression in *G. duodenalis,* or in the *BamHI/NotI* digested pQ30 vector (Qiagen, Germany) for expression in bacteria as N-terminal 6xHIS-tagged fusion protein.

Expression and purification of the recombinant proteins. Transformed *E. coli* were grown in SOB medium and recombinant proteins expression was induced at OD_{600}= 0.6, with 0.5 mM isopropyl-thio-β-D-galactoside at 37°C for 4 h. All 6xHIS-fused proteins were purified under native condition by affinity chromatography on nickel resin (Qiagen) and eluted with 250 mM imidazole (pH 8.0). Proteins were dialyzed against PBS (140 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4, pH 7.4) using Slide-A-Lyzer dialysis cassettes (cut-off 3.5 kDa, Thermo Fisher Scientific, Rockford, ILL, USA). The protein concentration was determined by Bradford assay (Thermo Fisher Scientific) and proteins were stored at –70°C until use.

Production of Polyclonal Antibodies Purified HIS-gG3PD fusion protein was used to immunize intraperitoneally two BALB/c mice (Charles River Laboratories International, Inc., MA, USA) on days 0, 21, and 42 with 50 μg protein in 300 μl of emulsified 1:1 PBS/Freund's complete adjuvant (Sigma Aldrich) (at day 0) or 1:1 PBS/Freund's incomplete adjuvant (Sigma Aldrich) (at day 21) or without any adjuvant (at day 42). Blood was collected from the tail vein before initial immunization and after each boost. Sera fractions were assayed for specific antibody content.

Preparation of* G. duodenalis *proteins. Total soluble proteins (S) and octyl β-D-glucopyranoside solubilized proteins from membranous material (M) were prepared as previously described (Lalle et al. 2012) starting from $2x10^9$ trophozoites or encysting parasites. The protein concentration was determined by Bradford methods (Thermo Fisher Scientific) and the protein lysates were stored at -70 °C. Alternatively, to assess the G3PD enzymatic activity in *G. duodenalis*, 10^6 parasites were collected, washed two times with cold PBS, suspended in 200 μl of PBS/1% TritonX100 and incubated 1 h on ice. After centrifugation at 13.000 x g for 15 min, supernatant was collected, protein concentration quantified by Bradford methods and the preparation was immediately used for the enzymatic assay as detailed below.

Western blot analysis. Proteins were separated on SDS-PAGE and transferred onto PVDF membrane with 39 mM glycine, 48 mM Tris, 0.1% SDS, and 10% methanol, using a semidry apparatus (BioRad, Hercules, CA, USA). Membranes were blocked with 5% skin milk in TTBS (20 mM Tris-HCl, pH 7.5, 100 mM NaCl, 0.05% Tween 20) for 1 h and then incubated with the primary antibody (Ab) in 2.5% skin milk/TTBS buffer. After incubation with an appropriate HRP-conjugated secondary Ab (1:3000), the interaction was revealed by chemiluminescence (Millipore, France). Antibodies were used at the following dilution: mouse polyclonal anti-gG3PD antiserum 1:3000; mouse anti-HA mAb (Sigma-Aldrich) 1:3000; rabbit N14 (anti-g14-3-3) antiserum (Lalle et al., 2006) 1:5000; mouse anti-αTubulin (Sigma-Aldrich) 1:10000; rabbit anti-gPGN (phosphoacetylglucosamine mutase, Lopez et al., 2003) 1:1000; mouse anti-HIS mAb (Qiagen) 1:2000.

Blue Native PAGE (BN-PAGE). 3–12% BN-PAGE (Invitrogen, Carlsbad, CA, USA) was carried out using 1 μg of purified HIS-tagged recombinant proteins according to the manufacturer. This technique allows the separation of very high molecular weight multiprotein complexes. Gels were run in Running buffer (0.002% Coomassie G-250, 50 mM BisTris/50 mM Tricine, pH 6.8) at 150 V for approximately 2 h. Gels were incubated in Tris/glycine/SDS buffer for 30 min and either stained with a Silver Staining kit (Invitrogen) or processed for western blot as described.

Affinity purification. FLAG-tagged proteins were purified by affinity chromatography on mouse anti-FLAG M2 mAb covalently bound to agarose beads (Sigma-Aldrich) as previously reported (Lalle et al., 2012) and directly eluted from the resin by incubation with 200 μM FLAG-peptide at 4 °C for 1 h. The eluted materials were stored at -70 °C until use.

Confocal Laser Scanning microscopy (CLSM). Trophozoites or encysting parasites were fixed and permeabilized as previously described (Lalle et al., 2006). Antibodies were used at the following dilution: polyclonal rabbit N14 antiserum 1:50; mouse polyclonal anti-gG3PD antiserum 1:50; polyclonal rabbit anti-gTom40 antiserum (Dagley et al., 2009) 1:50, FITC-conjugated mouse anti-HA mAb (Miltenyi Biotec, Germany) 1:50; Cy5-conjugated anti-CWP mAb (Waterborne Inc., New Orleans, LA, USA) 1:30. Alexa-Fluor 647- and 488-conjugated anti-rabbit and anti-mouse secondary Ab (Invitrogen) were used at a 1:500 dilution. After parasite staining, coverslips were extensively rinsed and then mounted using Vectashield® mounting medium (Vector Laboratories Inc., Burlingame, CA, USA) containing 300 nM of 4',6-diamidino-2-phenylindole (DAPI). CLSM observations were

performed on a Leica TCS SP2 AOBS (Leica Microsystems, Germany) apparatus, using excitation spectral laser lines at 405, 488 and 647nm and selecting emission wavelengths by a proper setting of the spectral detection system. Signals from different fluorescent probes were taken in sequential scan mode. Image acquisition and processing were conducted by using the Leica Confocal Software (Leica Microsystems). Image deconvolution was performed using the Huygens software (Scientific Volume Imaging BV, Hilversum, The Netherlands).

G3PD enzymatic assay, spectrophotometric and fluorimetric analysis. The dehydrogenase activity of gG3PD was assayed spectrophotometrically according to Kistler and Lin (1972) measuring the rate of phenazine methosulfate (PMS)-mediated reduction of the tetrazolium dye, MTT, to its formazan by addition of g3p. Reactions were performed in 96 well plate using either 100 µg of *G. duodenalis* Triton X-100 protein lysate or 16 pmol of purified recombinant HIS-tagged proteins in a final volume of 200 µl containing 67 mM of potassium phosphate (pH 7.5), 0.2% Triton X100, 6.6 µg of MTT, 20 µg of PMS, and 17 mM of g3p with or without 10 µM FAD. The g3p was omitted in the blank. Since crude extracts could contain substantial endogenous substrates, to ensure a linear rate of MTT reduction, reactions were pre-incubated at 25 °C for 5 min prior to the addition of MTT and g3p. Plates were sealed with an air-tight adhesive tape (Greiner Bio One, Austria) to prevent evaporation. After blank subtraction, the enzyme activity was calculated and expressed as micromoles per minute per milligram of protein, considering that the extinction coefficient of reduced MTT is 17 mM^{-1} cm^{-1} a 570 nm. UV-visible spectra of HIS-gG3PD (10 µg/µl) in 67 mM of potassium phosphate (pH 7.5), were recorded at 25 °C in 10 µl quartz capillaries. UV-visible spectra of 100 µM NBDHEX were recorded in 67 mM of potassium phosphate (pH 7.5), 17 mM of g3p, with or without HIS-gG3PD (42 pmol) in 0.5 ml quartz cuvette. All measures were done at 25 °C in a Multiskan Spectrum (Thermo) spectrophotometer. Fluorescent spectra (excitation at 430 nm) were acquired in a Luminescence Spectrometer LS50B (Perkin Elmer, Waltham, MA, USA).

Determination of intracellular glycerol. For the assay, 10^6 parasites were collected as previously described, suspended in 100 µl of PBS, incubated 10 min at 95°C and the supernatant collected after centrifugation at 12,000 x g for 1 min. Protein concentration was measured by Bradford assay. Intracellular glycerol level was enzymatically determined using the Free Glycerol Reagent kit (Sigma-Aldrich) by measuring spectrophotometrically the production of a quinoneine dye at 540 nm. Assay was performed on 96 well-plate according to the manufacturer's instruction, using 40 µl of parasite

lysate and 160 µl of enzyme mixture per well. After 5 min of incubation at 37 °C, absorbance was read by Multiskan Spectrum (Thermo Scientific) spectrophotometer. Intracellular glycerol concentration was calculated interpolating the obtained absorbance with a glycerol standard curve and expressed as pmol of glycerol/µg of total protein. Each condition was assayed in triplicate and the experiment was performed independently three times.

In vitro drug susceptibility assay. The *in vitro* assays were performed according to Bénéré et al. (2007) and Hounkong et al. (2011) with modifications. All compounds were dissolved and serially diluted in 1:1vol ethanol/DMSO. *G. duodenalis* trophozoites (5×10^5 parasite/ml) were cultured in 96-well plates (Nunc Δ surface, Thermo Scientific) in TYIS-33 medium and 3 µl of 100X concentrated compound, or solvent, were added to reach the desired compound concentration. The plates were sealed with an air-tight adhesive tape (Greiner Bio-One GmbH, Austria) and incubated for 48h at 37°C. After the incubation period, culture medium was removed. Adherent trophozoites were immediately fixed with 300 µl of methanol for 2 min and then stained with a solution of 0.1% methylene blue in PBS for 10 min at RT. Wells were washed 3 times with PBS and incubated with 200 µl of ethanol and 0.1 M HCl (1:1 (v/v) in H_2O), to elute the dye. Absorbance was determined at 655 nm by Multiskan Spectrum (Thermo Scientific) microplate spectrophotometer. Growth inhibition was calculated as the percentage of treated parasite in comparison with untreated parasite. Three independent experiments were performed and each drug dilution was assayed in triplicate. Alternatively, for drug efficacy tests with resazurin (Bénéré et al. 2007), trophozoites were grown in the presence of serial dilutions of the drugs, or DMSO as control, in an anaerobic growth chamber (100 % N_2, 37°C). After 72 h, the medium was removed, the wells were washed three times with pre-warmed PBS containing 1% of glucose (0.2 ml per well), and finally, 0,2 ml PBS with glucose containing 10 mg/l resazurin were added. Reduced resazurin was quantified by fluorimetry (excitation at 365 nm, emission at 455 nm) using a 96-well-multimode plate reader Enspire (Perkin-Elmer).

For immunolocalization and enzymatic assays, drug treatments were performed in 10 ml screw cap tubes using $1x10^5$ parasite/ml of *G. duodenalis* trophozoites. Twenty five µl of 20 mM NBDHEX in ethanol, or the equal volume of ethanol, was added to 10ml of TYS-I-33 medium (final concentration 50 µM) and parasites were incubated at 37°C for the indicated time periods and cells were processed as described in previous paragraphs.

Mass spectrometry analysis. Proteins were separated on a 1D-gel NuPAGE 4-12% (Novex, Invitrogen) run in morpholinepropanesulfoninic acid (MOPS) buffer and stained with the Colloidal Blue Staining kit (Invitrogen). Slices were excised and digested with modified sequencing-grade trypsin (Promega Corporation, France), as previously described (Shevchenko et al., 1996). Nanoflow reversed-phase liquid chromatography tandem mass spectrometry (RP-LC-MS/MS) analysis of peptide mixtures was performed using an HPLC Ultimate 3000 (DIONEX, Sunnyvale, CA U.S.A) coupled with a linear ion trap (LTQ, Thermo Electron, San Jose, CA) mass spectrometer. Peptides were desalted in a trap column (Acclaim PepMap 100 C18, LC Packings, DIONEX) and separated in the analytical column, a 10 cm long fused silica capillary (Silica Tips FS 360-75-8, New Objective) in house slurry-packed with a 5 μm, 200 Å pore size C18 resin (Michrom BioResources, CA). Peptides were eluted using a 40 min linear gradient from 96% aqueous buffer containing 5% acetonitrile and 0.1% formic acid to 60% organic buffer constituted by acetonitrile with 5% H_2O and 0.1% formic acid, at 300 nL/min flow rate. Analyses were performed in positive ion mode and the HV Potential was set up around 1.7-1.8kV. Tandem mass spectra were matched against the *G. duodenalis* protein database (Giardia DB version 1.2) downloaded from the Web site http://www.giardiadb.org/giardiadb and through the SEQUEST algorithm incorporated in the Bioworks software (version 3.3, Thermo Electron). The following match parameters were considered: fully tryptic cleavage constraints (one miscleavage allowed), static cysteine carbamidomethylation, variable methionine oxidation. Precursor and fragment ions were searched with 1.5 and 1 Da tolerance, respectively. Possible NBDHEX adducts on cysteine or lysine residues were searched. The mass increments considered were +296, +281, +265 Da, for the intact, partially or completely nitro-reduced NBDHEX adducts, respectively. Statistical parameters used for legitimate protein identification were described elsewhere (Lalle et al., 2012).

Chemical reduction of NBDHEX. Sodium dithionite was prepared as a 1M solution in ddH_2O and was used 100 mM to chemically reduce 10 mM NBDHEX in 20 µl 0.5% NH_3. After 5 min, 2 µl reaction was diluted to 200 µl in 67 mM potassium phosphate buffer (pH 7.5). NBDHEX fluorescence quenching was verified at 25°C in a Luminescence Spectrometer LS50B (Perkin Elmer) whereas the UV-visible spectra was acquired at 25°C in a Multiskan Spectrum (Thermo Scientific) spectrophotometer in 0.5 ml quartz cuvettes. For the mass spectrometry analysis, the reaction mixture was diluted in 50% ethanol and 1% ammonia and directly infused in the LTQ mass spectrometer through a glass tip. The nanospray ionization in positive ion mode was allowed applying 1.5-1.6kV.

HV Potential. Full MS in the 150-350 m/z range were acquired; then MS3 fragmentation of the ions 280 and 250, derived from MS2 of 298 and 268, respectively, was induced.

Sequence analysis. Conserved functional domains and sites in the protein sequence were search using ELM (Dinkel et al., 2014) and BLASTP (http://blast.ncbi.nlm.nih.gov) algorithms. Transmembrane regions were search by TMHMM Server v.2.0 (http://www.cbs.dtu.dk/services/TMHMM/). Mitochondrial targeting signal were search using Psort II (http://psort.hgc.jp/form2.html) and MitoProt II (http://ihg.gsf.de/ihg/mitoprot.html). Genomic sequences of 1133 organisms were downloaded from EggNog database v.3.0 (Powell et al., 2012). If more than one strain was available, only the strain classified as "core" was retained. This selection resulted in 774 organisms, which were included in the reference set. The Smith-Waterman algorithm available in the FASTA package (Pearson and Lipman, 1988) was exploited using as query the gG3PD full-length sequence, or the protein portions encompassing: the first 500 residues (referred as N-terminal domain); or the residues 501-950 (referred as central domain); or the residues 951-1111 (referred as the C-terminal domain); or the protein portion including the N- and central domains (residues 1-950). Sequences having a match result with E-value lower than 10^{-6} were selected, except for the C-terminal domain for which the E-value threshold was set at 10^{-3}. Sequence analysis, editing and phylogenetic analysis were performed using Bioedit (Hall T.A., 1999), MEGA 5.0 (Tamura et al., 2013) and Jalview 2.8.2 (Waterhouse et al., 2009).

Statistics. Statistical analyses were performed with PRISM 6.0 software (GraphPad Software, Inc; La Jolla, CA, USA) and significance was calculated by Student unpaired t-test and one-way ANOVA. A P-value <0.05 was considered statistically relevant.

Results

Sequence analysis of the putative FAD-dependent glycerol-3-phosphate dehydrogenase (gG3PD) of G. duodenalis reveals an unusual protein organization. The GL50803_16125 entry encodes a protein of 1111 amino acids (approximately 119 kDa) annotated as FAD-dependent oxidoreductase/glycerol-3-phosphate dehydrogenase, that we named gG3PD. A combined search for conserved regions in the gG3PD protein sequence retrieved three principal domains (Supplemental Figure 1A). The region encompassing residues 31-510 contains the multidomain TIGR03377 corresponding to the protein family of the subunit A (GlpA) of anaerobic GlpACB. A second domain, residues 587-910,

corresponds to a small NADH binding domain within a larger FAD binding domain (Pyr_redox_2, PF07992) which is common to flavoproteins of the pyridine nucleotide-disulphide oxidoreductases family (e.g. thioredoxin reductases, NADH oxidases and peroxidases). The last domain (DUF 1667), encompassing residues 1039-1103, was previously found in archaeal and bacterial hypothetical proteins, some of which are annotated as being potential metal-binding proteins. As for other G3PDs, no transmembrane (TM) regions could be detected. Due to its unusual domain organization, the full-length or three portions (residues 1-510, 511-950 and 951-111) of the gG3PD protein sequence were used as query against a large non-redundant set of high quality reference genomes obtained from EggNog. Despite an homologous sequence present in the genome of the closely related diplomonad *Spironucleus salmonicida*, only protozoan parasites of the genus *Entamoeba* encode for a single sequence orthologous to the full length gG3PD and also annotated as glycerol-3-phosphate dehydrogenase. One orthologue of the first two portions of the gG3PD was found in the genomes of *Eggerthella* spp., a non-sporulating medically important anaerobic Gram-positive bacillus, and annotated as FAD-dependent oxidoreductase. On the contrary several hundred orthologues of the first portion of gG3PD (residues 1-510), all annotated as glycerol-3-phosphate-dehydrogenases, were found in the genome of human and animal pathogenic bacteria, such as *Trepomena* spp. (Spirochaetes) and *Clostridium* spp. (Firmicutes), or to halophilic archaea, e.g. *Halorhabdus utahensis,* (Supplemental Figure 1B). More than one hundred orthologues of the second portion (residues 501-950) were found in human and animal pathogenic bacteria, but also in soil and plant bacteria (data not shown). Although several of these proteins were annotated as oxidoreductases it was not possible to define, in terms of substrate or biochemical pathway, any specific enzymatic function. Finally, only 33 orthologues were found respect to the third portion of gG3PD (residues 951-111), which putative function could be that of metal-binding proteins.

The intracellular localization and stage expression of gG3PD protein. According to the data reported in GiardiaDB, the expression level of the *gg3pd* is downregulated during encystation. To investigate whether the gG3PD protein expression level was similarly regulated we used a mouse polyclonal antibody raised against a full-length HIS-tagged recombinant gG3PD expressed in bacteria. As shown in Figure 1A, a band of approximately 120 kDa was immunodecorated with comparable intensity by the anti-gG3PD Ab, both in cellular lysates from trophozoites and from parasites harvested at different time points after encystation induction (at 6h, 12h and 24h). Encystation progress was checked by immunostaining with anti-phosphoacetylglucosamine mutase (gPGN) Ab, a protein induced during

365 cyst formation (Lopez et al., 2003). The anti-gG3PD Ab was also used to study the intracellular
366 localization of the protein by confocal laser scanning microscopy (CLSM). In the trophozoite (Figure
367 1B, panel a) the anti-gG3PD strongly stained the plasma membrane of the parasite ventral surface
368 likely corresponding to the marginal groove and ventrolateral flanges (VLF), fingerlike projections of
369 the plasma membrane involved in parasite attachment to the surface and to the host cell (Sousa et al.,
370 2001). As previously reported, the g14-3-3 showed a spotty broad distribution in the parasite cell body,
371 excluding the nuclei and the median body (smile-like microtubule aggregate in the middle of parasite
372 body). The widespread g14-3-3 staining makes it difficult to claim for a discrete co-localization with
373 the gG3PD, although a slight co-localization signal (pseudo color yellow) was visible at the posterior
374 edge of the trophozoite (Figure 1B, panel a, merge). *G. duodenalis* encystation was monitored by the
375 progressive appearance of specialized structures for the secretion of the cyst wall proteins, the
376 encysting specific vesicles (ESVs), which were stained, together with the cyst wall, by the anti-cyst
377 wall protein (CWP) mAb. In encysting parasites, in addition to the ventral plasma membrane, the anti-
378 gG3PD Ab (Figure 1B, panel b and c) labeled several structures, similar to aggregates/small vesicles, in
379 the cytoplasm. These structures did not co-localize with the ESVs and no straightforward co-
380 localization with g14-3-3 could be detected (Figure 1B, panel b and c). In cysts, the gG3PD was
381 diffused in the cytoplasm, although a marked aggregation close to the plasma membrane, below the
382 cyst wall was evident (Figure 1B, panel d). A faint co-localization with g14-3-3 (pseudo color yellow)
383 was detected below the cyst wall.
384 Since the staining with the anti-gG3PD Ab observed in encysting parasites resembled that of
385 mitosomes, additional co-localization experiments were performed using an antibody directed against
386 the mitosomal marker Tom40 (Dagley et al., 2009). As shown in Figure 1C, both in trophozoites and
387 encysting parasites (panel a and b), the anti-Tom40 Ab stained the main central mitosome between the
388 nuclei and several peripheral smaller mitosomes. In trophozoites no co-localization with anti-G3PD
389 could be detected (Figure 1C, panel a), whereas in encysting parasites superimposable signals (pseudo
390 color yellow) were evident in some of the peripheral mitosomes, and to a lesser extent in the central
391 one (Figure 1C, panel b), suggesting a partial re-localization of the protein to mitosomes.
392
393 ***Interaction between gG3PD and g14-3-3.*** To further characterize the gG3PD, a N-terminally FLAG-
394 HA tagged gG3PD was expressed in *G. duodenalis* under the *α-tubulin* constitutive promoter. The
395 FLAG-HA-gG3PD was expressed at the same level, and comparable to the endogenous protein, both in
396 trophozoites and during the encystation stage (Figure 2A) and the intracellular localization largely

resembled that of the endogenous gG3PD (Supplemental Figure S2). By following protein fractionation (soluble and membrane fractions), the endogenous and the FLAG-HA-gG3PD were mainly detected in the soluble fraction at both stages (Figure 2B). Whereas, a small protein amount could be detected in the membrane fraction only at the trophozoite stage (Figure 2B), partially in agreement with immunolocalization observations, suggesting that the gG3PD could form strong interactions with the membrane or with membrane proteins. To better investigate the previously suggested interaction between g14-3-3 and gG3PD (Lalle et al., 2012), immunoprecipitation experiments were performed. As shown by anti-HA immunoblotting (Figure 2C, upper panel), the anti-FLAG mAb immunoprecipitated the FLAG-HA-gG3PD only from transfected parasite extracts. Despite FLAG-HA-gG3PD was comparably immunoprecipitated from both trophozoites and encysting parasites, g14-3-3 was mainly co-immunoprecipitated from trophozoites, whereas only a faint signal was observed in the immunoprecipitate from encysting parasites (Figure 2C, lower panel).

To shed light on the relevance of the two principal domains characterizing the gG3PD protein, two deletion mutants were constructed and independently expressed in *G. duodenalis* as FLAG-HA-tagged proteins. The FLAG-HA-gG3PD_N corresponds to N-terminal half of the protein (Supplemental Figure S1A) and contains the FAD-dependent anaerobic glycerol-3-phosphate dehydrogenase-like domain (residues 2-530); whereas the FLAG-HA-gG3PD_C corresponds to the C-terminal half and contains the FAD-dependent pyridine nucleotide-disulphide oxidoreductase-like domain (residues 524-1111). As shown by immunoblot (Figure 3A, upper and middle panels), the expression of both FLAG-HA-tagged proteins was detected at both stages. In all transfection experiments, the expression of deletion mutants was lower than that of the full-length FLAG-HA-gG3PD.

Intracellular localization of both mutants was performed with anti-HA antibody. CLSM analyses (Figure 3B, panels a and b) showed that, in trophozoites and encysting parasites, the FLAG-HA-gG3PD_N partially localized to the ventral plasma membrane and ventrolateral flanges, like the endogenous gG3PD. Although a spotty cytoplasmic and an intense nuclear/perinuclear staining were evident, in contrast with the endogenous and the full-length FLAG-HA-gG3PD, suggesting that a fraction of the deleted protein could be misfolded and/or mislocalized. In FLAG-HA-gG3PD_C transgenic parasites the protein was completely mislocalized and appeared widespread distributed, also inside nuclei, with the nuclear/perinuclear localization even more evident in encysting parasites (Figure 4B, panel d and e). No mitosome-like staining was observed in encysting parasites for any of the mutants (Figure 3B, panels b and e). Few cysts with a spotty cytoplasmic localization of the protein were observed in both transgenic strains (Figure 3B, panels c and f). These observations likely suggest,

that the N-terminal half of the protein contains the structural/sequence determiners necessary to localize the gG3PD at the ventral plasma membrane and ventrolateral flanges. Looking for the domain(s) involved in the interaction with g14-3-3, co-immunoprecipitation with anti-FLAG was also performed using the FLAG-HA deletion mutants. Despite several attempts no conclusive results were obtained (data not shown), likely due to the low expression of these proteins or to the presence of other factors that could affect the interaction (e.g. partial misfolding, incomplete phosphorylation).

Evaluation of gG3PD enzymatic activity. To better define the role of gG3PD during *G. duodenalis* differentiation, we first characterized the protein activity using N-terminally HIS-tagged versions of both full-length gG3PD and deletion mutants heterologously expressed in *E. coli*. All recombinants were expressed as soluble proteins and could be purified (up to 95%) under native conditions (Supplemental Figure S3). The purified full-length HIS-gG3PD, but not the deletion mutants, showed a distinctive yellow color (data not shown), compatible with the presence of a flavin cofactor. Indeed, the UV-visible spectra of the HIS-gG3PD showed two absorbance peaks with a maximum approximately at 350 and 440 nm (Figure 4A), very close to those of free flavin (either FAD or FMN or both), supporting that the gG3PD is a flavoprotein. No similar peaks could be observed in the spectra of any deletion mutants (data not shown). Since mG3PD/glpD are homodimeric proteins (Mráček et al., 2013), the dimeric nature of gG3PD was assayed using native PAGE. As shown in Figure 4B, the HIS-gG3PD_N was exclusively monomeric (60 kDa); whereas the HIS-gG3PD migrated in the gel both as a 120 kDa monomer and as a 240 kDa dimer, as confirmed by immunoblot assay with anti-HIS. Similarly, the HIS-gG3PD_C migrated as a monomer around 60 kDa, and as a dimer of 120 kDa, thus suggesting that at least one function of the gG3PD C-terminal half is to mediate protein dimerization.

The FAD-glycerol-3-phosphate dehydrogenase activity of the HIS-gG3PD was then assayed by measuring the g3p-dependent reduction of MTT in presence of PMS, which mediates the transfer of reducing equivalent from the enzyme to the terminal dye (Kistler and Lin, 1972). As shown in Figure 4C, the HIS-gG3PD displayed a specific activity corresponding to 86.7 ± 5.6 nmol/min/mg in the absence of exogenous FAD, that did not significantly increase (89.8 ± 0.7 nmol/min/mg) even when 10µM FAD was added. Moreover, no activity could be observed either in the absence of g3p or PMS (data not shown), thus proving that g3p is indeed the substrate of the enzyme and that an electron carrier molecule, such as PMS, is required to reduce MTT. This is in agreement with the role of quinone described for other FAD-dependent G3PDs (Unden and Bongaerts, 1997; Mráček et al., 2013). Intriguingly (Figure 4C), the HIS-gG3PD_N, containing the GlpA-like domain, also displayed a basal

461 enzymatic activity (30.5 ± 1.0 nmol/min/mg), 2.8-fold lower in comparison with the full length protein,
462 and it showed an increase of 17% in the activity (36.7 ± 3.0 nmol/min/mg) after stimulation with
463 exogenous FAD. On the contrary, the HIS-gG3PD_C did not show any enzymatic activity in any
464 condition (data not shown). These results support the homology data and confirm that the glycerol-3-
465 phospate dehydrogenase activity is localized to the N-terminal half of the gG3PD.
466 To study the possible correlation between encystation-dependent intracellular re-localization of gG3PD
467 and its function, the enzyme activity was measured in trophozoites and encysting parasites extracts.
468 Indeed, compared to trophozoites, the gG3PD activity showed a 20% increase at 6h of encystation
469 (Figure 4D), then it was reduced nearly down to the trophozoite level after 12h of encystation. No
470 variation in the expression level of the gG3PD protein could be observed at any time points (Figure
471 4E). The encystation process in *G. duodenalis* has been proposed as a primitive response to cellular
472 stress (Argüello-Garcia et al., 2009). In unicellular organisms, as response to stress signals, the
473 intracellular level of glycerol increases, as consequence of a massive conversion of DAPH to g3p, due
474 to the greater activity of the NADPH-dependent G3PDs and in some cases also to that of FAD-
475 dependent G3PDs (Yang et al., 2007). To test this hypothesis we measured the intracellular content of
476 glycerol during encystation. The amount of glycerol increased up to 60% already at 6h post-encystation
477 induction (Figure 4F) and then of an additional 15% in 12h encysting parasites, thus suggesting a
478 possible relationship between the enhanced gG3PD activity and the glycerol accumulation occurring
479 during parasite differentiation.
480
481 ***In vivo effects of NBDHEX on G. duodenalis survival and gG3PD activity.*** Recently, a 7-nitro
482 benzoxadiazole (NBD) derivative has been shown to impair the growth of PC-3 adenocarcinoma cell
483 line by a significant inhibition of the glycerol-3-phosphate oxidoreductase activity of mG3PD (Singh,
484 2014). We wondered whether this class of molecules could be effective against *G. duodenalis*
485 trophozoites and the gG3PD. We selected the 6-(7-nitro-2,1,3-benzoxadiazol-4-ylthio)hexanol
486 (NBDHEX), a promising antitumoral drug acting as a suicide inhibitor of human glutathione S-
487 transferases (GSTs), that also shows a good tolerance in mouse models (Ricci et al., 2005; Turella et
488 al., 2005; Pellizzari Tregno et al., 2009; Tentori et al., 2011). NBDHEX cytotoxicity against *G.*
489 *duodenalis* trophozoites was evaluated *in vitro* after 48h of treatment in microaerophilic growth
490 conditions, using a colorimetric assay. As shown (Figure 5A), NBDHEX was effective against the
491 parasite (IC_{50}: 0.3 ± 0.1 µM) at a lower concentration (5.6-fold) than that of the reference drug MTZ
492 (IC_{50}: 1.5 ± 0.1 µM). Moreover, under anaerobic growth conditions NBDHEX was also more effective

than MTZ (IC_{50}: 0.6 ± 0.4 µM for NBDHEX and IC_{50}: 1.9 ± 0.2 µM for MTZ) (data not shown). To maximize in a short time period the possible NBDHEX effects, taking into account the parasite intracellular environment, trophozoites were treated *in vivo* with 50 µM NBDHEX and then the gG3PD enzymatic activity was assayed in parasite extracts. A 50% decrease of the gG3PD activity was already evident after 2h of treatment (Figure 5B, $P< 0.01$), and a further 5% decrease was achieved after 4h and 6h. This suggests that *in vivo* NBDHEX treatment reduces the gG3PD activity, without inducing any statistically relevant variation of the protein expression level (Figure 5C and D).

Interaction of NBDHEX with gG3PD. To ascertain if the effects on the gG3PD activity were due to a direct NBDHEX binding to the enzyme, co-localization experiments were performed by CLSM (Figure 6A), taking advantage of the fluorescent properties of the compound (maximum emission peak at 525 nm) (Ricci et al., 2005). Only trophozoites treated with NBDHEX showed a spotted faint fluorescence in the cytoplasm. A more intense signal was observed around the nuclei, in the areas corresponding to the median body, in the intracellular portion of ventral flagella and, intriguingly, at the ventral plasma membrane. As shown, co-localization signal (pseudo color yellow) between gG3PD and NBDHEX was evident only at plasma membrane, without any relevant alteration in the intracellular localization of gG3PD (for comparison see Figure 1B and C, panels a). To further study the interaction between NBDHEX and gG3PD, *E. coli* expressing the HIS-gG3PD was incubated with NBDHEX, then the recombinant protein was purified and, finally, the enzymatic activity was assayed. Similar to the *in vivo* treatment of *G. duodenalis* with NBDHEX, incubation of bacteria with the compound resulted in a recombinant HIS-gG3PD having a 25% reduced enzymatic activity (Figure 6B). Next, we wondered whether NBDHEX could bind HIS-gG3PD. We take advantage of the property of NBD derivatives covalently bound to a protein to be visualized by fluorescence under ultraviolet (UV) light after SDS-PAGE (Bragg and Hou, 1999). Indeed, HIS-gG3PD, purified from NBDHEX treated bacteria, was fluorescent under UV-light (Figure 6C, right panel), even in the presence of a reducing agent (such as DTT). On the contrary, no fluorescence was associated with the HIS-gG3PD purified from untreated bacteria. These results suggest that NBDHEX strongly binds HIS-gG3PD. To confirm this hypothesis, purified HIS-gG3PD from bacteria, treated or not with NBDHEX, was subjected to mass spectrometry analysis. MS/MS spectra were acquired matching two HIS-gG3PD peptides carrying NBDHEX derived adducts on cysteine residues: the detected mass shift was compatible with an NBDHEX form in which the nitro group was reduced to amine (Figure 6D). No adduct with the intact NBDHEX moiety could be found, probably because it was unstable or not detectable under the applied

experimental conditions. The presence of nitro-reduced NBDHEX adducts was in favor of an electrochemical reduction of the drug, likely by gG3PD. Therefore, NBDHEX properties were evaluated by spectrophotometric and fluorimetric analyses after *in vitro* incubation with HIS-gG3PD. The UV-visible spectra of NBDHEX, incubated either with g3p or with HIS-gG3PD, showed the typical peak centered around 430 nm (Figure 7A), even after 80 min of incubation (data not shown). Incubation in the presence of both HIS-gG3PD and g3p led to the progressive decrease of the 430 nm absorption peak and to the appearance of a less intense new one around 450-455 nm (Figure 7A). The reaction was also associated with a change in color, from bright yellow to brown (Figure 7B, insert) and the disappearance of the 525 nm emission peak of NBDHEX in the fluorescence spectra (Figure 7B). Since the amount of NBDHEX was in large excess compared to HIS-gG3PD, the observed spectral alterations were not imputable to binding of NBDHEX to the enzyme, but they argued on NBDHEX modification by the glycerol-3-phosphate dehydrogenase activity of HIS-gG3PD. Such modifications could be ascribable to a partial or complete reduction of the nitro group to hydroxylamine or amine. Chemical reduction of the NBDHEX nitro group to amine by sodium dithionite, proved by mass spectrometry analysis (Supplemental Figure S4), resulted in the decrease of the NBDHEX absorption peak at 430 nm, although with the appearance of a new peak at 405 nm (Figure 7C), and loss of fluorescence too (Figure 7D). The observed differences between UV-visible spectra of HIS-gG3PD and sodium dithionite treated NBDHEX suggest that the enzymatic activity produced an optical species, probably compatible with an incomplete reduced form of NBDHEX, carrying an hydroxylamine instead of an amine moiety. On the other side, looking to the NBDHEX-treated HIS-gG3PD, both in the presence or absence of g3p, the enzyme resulted fluorescent under UV light after SDS-PAGE (Figure 7E). This suggests that protein adducts with the unreduced NBDHEX occurred despite, enzyme activation. Products from *in vitro* reactions between HIS-gG3PD and NBDHEX were also analyzed by mass spectrometry. In the sample incubated in the presence of g3p, some HIS-gG3PD cysteine residues were modified by NBDHEX and MS/MS spectra analysis allowed to detect mass shifts compatible with NBDHEX adducts carrying the NO_2 moiety completely reduced to NH_2, but also the semi-reduced idroxylamine group (Supplemental Figure S5). These adducts could not be detected when g3p was omitted.

Since toxicity of the nitrocompound MTZ over *G. duodenalis* has been related to its reduction to nitrosoimidazole or hydroxylamine intermediate by flavin enzymes (Leitsch et al., 2007; Leitsch et al., 2009; Leitsch et al., 2011), we asked whether the gG3PD could be able to nitroreduce MTZ. We then monitored variation in the MTZ absorbance at 320 nm (Leitsch et al., 2011). As shown, no decrease in

MTZ absorbance occurred even after 80 min of incubation in the presence of HIS-gG3PD and g3p (Figure 7F), thus suggesting that in our experimental conditions MTZ is not nitroreduced by gG3PD.

Discussion

In this work we have characterized, for the first time, the FAD-dependent glycerol-3-phosphate dehydrogenase of *G. duodenalis* both at molecular and functional level, and we have shown that the antitumoral 7-nitro benzoxadiazole derivative NBDHEX displays a remarkable antigiardial activity, targets the gG3PD and, when *in vivo* administered to *G. duodenalis*, induces gG3PD activity reduction.

The full length gG3PD has no orthologues in other organisms, including humans, except for the diplomonad *S. salmonicida* and the amebidae *Entamoeba* spp. The presence in the gram-positive anaerobic actinobacteria *Eggerthella* spp. of an orthologue of the first 950 amino acids suggests that a first gene fusion event between a GlpA-like glycerol-3-phosfate-dehydrogenase and a pyridine nucleotide-disulphide oxidoreductase occurred in this prokaryotic lineage. A lateral gene transfer (LGT) event could account for the presence of this multi-domain protein both in *G. duodenalis* and *Entamoeba* spp., that share with *Eggerthella* spp. the mammalian intestine as ecological niche. LGT events have been already well documented for genes of the anaerobic metabolism in eukaryotic protozoan parasites such as *E. hystolitica, T. vaginalis,* and *G. duodenalis* (as well as in *S. salmonicida)* (Andersson et al., 2007; Morrison et al., 2007; Alsmark et al., 2009).

We demonstrate that the gG3PD is an active flavoenzyme able to use g3p as source of reducing equivalents. In agreement with the domain topology and homology with bacterial GlpA subunit, the enzymatic activity, as proven by deletion mutants, resides in the N-terminal half of the protein. It is worth noting that the modest stimulatory effect exerted by additional FAD on the enzymatic activity of both the gG3PD and the gG3PD_N, could be explained by different reasons i) the full length gG3PD is fully complexed with FAD; and/or ii) the deletion mutant lacks the ability to properly bind FAD; and/or iii) the enzyme needs FMN in addition to FAD. Indeed, the maximal activity of *E. coli* trimeric GlpACB *in vitro* requires both FMN and FAD (Cole et al., 1988), whereas glpD binds only FAD (Yeh et al., 2008). Furthermore, we show that gG3PD is able to dimerize and that the C-terminal half is required for the dimer formation. The dimerization properties of this gG3PD region, absent in bacterial GlpA, could be ascribable to its homology with members of the FAD-dependent pyridine nucleotide-disulphide oxidoreductases, a family of generally dimeric proteins (Argyrou and Blanchard, 2004).

Localization of the gG3PD to the ventral plasma membrane, likely in correspondence to the ventrolateral flanges, and the detection of the enzymatic activity in parasite extracts are in favor of the

existence of functional glycerol biosynthetic pathway and of a role of the protein in *G. duodenalis* energy metabolism. Similar to bacteria (Unden and Bongaerts, 1997), metabolization of g3p by *G. duodenalis* could occur at plasma membrane via gG3PD thus providing an extra source of DHAP for glycolysis and reducing equivalents to putative ETCs. Experimental evidences indicate that *G. duodenalis* can proliferate *in vitro* even in absence of glucose, its main energy source (Adam, 2001), either in the presence of serum (Schofield et al., 1991) or of a defined mixture of bile salts, phosphatidylcholine (PC) and cholesterol (Gillin et al., 1986; Gault et al., 1987). Incorporation of free glycerol and g3p in trophozoites has been reported to be negligible (Jarroll et al., 1981). Neverthless, mammalian bile and intestinal epithelium mucus are rich of PC, that is efficiently taken up by *G. duodenalis* trophozoites (Stevens et al., 1997). PC, in turns, can potentially be converted to g3p by the combined activity of a phospholipase PLB and a putative GPC-PDE, both encoded by *G. duodenalis* genome (Yichoy et al., 2011; Morrison et al., 2007).

We show that the gG3PD activity increases early at encystation, without up-regulation of protein expression. In this phase of encystation, energy metabolism and galactosamine synthesis require extra carbon sources as consequence of the highly reduced glucose uptake and decrease in oxygen consumption (Paget et al., 1998). We may speculate that gG3PD could provide extra carbon to glycolysis, via an improved metabolization of g3p to DHAP. In parallel, we also observed an increase in intracellular glycerol levels. In response to stress conditions, including hyperosmotic stress, several microorganisms synthesized glycerol as stress protector via an increased DHAP to g3p conversion mediated by the NAD(P)H-dependent G3PD (Albertyn et al., 1994; Yang et al., 2007; Chen et al., 2012; Suescún-Bolívar and Thomé, 2015). The increase in the bile salt, together with a slight alkaline pH, is a key encystation signal eventually perceived by *G. duodenalis* as hyperosmotic stress. It is worth to note that, in *E. hystolitica,* intracellular accumulation of both g3p and glycerol occurs in response to oxidative stress despite only the gG3PD orthologue is present and no NAD(P)H-dependent G3PD activity has been detected (Husain et al., 2012). The unusual domain architecture of FAD-dependent G3PDs from both *G. duodenalis* and *E. hystolitica* may explaind for a bi-directional activity of this enzymes A further characterization of the gG3PD and a detailed metabolomics analysis of the trophozoite and encystation stage of *G. duodenalis* will be necessary to unravel any metabolic re-arrangement occurring during the parasite differentiation and its link with gG3PD activity.

The subcellular localization of gG3PD requires also other considerations. The presence, only in trophozoites but not in encysting parasites, of a fraction of the enzyme in the membrane preparation is in agreement with the immunolocalization data and suggests that gG3PD can associate with

membranes in a stage dependent-manner. The structural/sequence determinant(s) responsible for the localization of the protein may reside in the N-terminal half of gG3PD, as supported by the partial localization of the gG3PD_N deletion mutant to the ventrolateral side of the trophozoite. In addition, we confirm that gG3PD and g14-3-3 co-precipitate (this work; Lalle et al., 2012) mainly at the trophozoite stage implying a role of the g14-3-3 in gG3PD localization to ventral plasma membrane. The presence of several putative mode-1 and mode-2 binding sites along the sequence of gG3PD (Supplemental Figure S1; Lalle et al., 2012) can allow for a direct interaction between the two proteins. Nevertheless, we cannot exclude that the interaction between gG3PD and g14-3-3 is indirect and that both proteins are present in a multiprotein complex, and further studies with different deletion and point mutants will clarify this issue. However, beyond the exact interaction mechanism, a clear relationship between g14-3-3 and gG3PD localization and, possibly, its enzymatic activity, is indicated by the strongly reduced co-precipitation of g14-3-3 with gG3PD during encystation and the partial re-localization of gG3PD to mitosomes.

The re-localization of gG3PD to mitosomes, despite the absence of any predictable mitochondrial targeting sequence or cleavage site in the N-terminus, has been proved at least for another *G. duodenalis* protein, namely GiiscS (Dolezal et al., 2005). In our case, the lost of mitosomal localization in both gG3PD deletion mutants suggests that targeting to these organelles at least requires the full-length protein. Although *G. duodenalis* mitosomes clearly miss the TCA cycle and the typical ATP-generating mitochondrial respiratory ETC (Han and Collins, 2012), the encystation-dependent localization of gG3PD to mitosomes could be a reminiscence of mitochondrial function of mG3PDs, where they participate in the g3p shuttle and supply electrons to the mitochondria ETC. A homolog of FAD-dependent mG3PD has been identified also in the mitosomes of the microsporidia *Antonospora locustae*, a group of spore-forming fungus-related intracellular parasites (Dolgikh et al., 2011). The protein is expressed in the spore and localizes in the mitosomes, where it channels reducing equivalent to the so called alternative oxidase (Dolgikh et al., 2011). Similarly, the presence of alternative ETCs in *G. duodenalis*, as in the anaerobic bacteria respiration, cannot be excluded (Unden and Bongaerts, 1997; Han and Collins, 2012). GiOR-1, a di-flavoprotein able to reduce the gCYTb5-IV and having ferredoxin-independent and NADPH-dependent reductase activity, has been localized to the *G. duodenalis* mitosome, suggesting that a NAD(P)-dependent ETC exists in this organelle (Jedelský et al., 2011; Pyrih et al., 2014). Similarly, the presence of a pyridine nucleotide-disulphide oxidoreductase domain and of a putative metal binding domain in gG3PD, both present in the flavoprotein reductase family (Argyrou and Blanchard, 2004), may imply that the protein possesses a combined

dehydrogenase and reductase activity, although the final acceptor, if any, is still to be identified. Our results disclose a new perspective on the role of mitosome in *G. duodenalis* suggesting that mitosomal components and functions can change during the parasite life cycle. Indeed, gG3PD was not detected in the proteome of mitosomes isolated from trophozoites (Jedelský et al., 2011).

Looking for compounds affecting the gG3PD activity, we demonstrate that the 7-nitro benzoxadiazole derivative NBDHEX, a patented anti-tumor drug (Ricci et al. 2005; Turella et al. 2005), exerts cytotoxic activity towards *G. duodenalis* trophozoites and, when administered *in vivo,* hampers the g3p dehydrogenase activity of gG3PD. The IC_{50} of NBDHEX is 2- to 5-fold lower than MTZ, proving that the compound is more effective in killing *G. duodenalis* than the reference drug, even in microaerobic conditions. This indicates that the presence of O_2 does not produce opposite effects on NBDHEX, as on the contrary occurs for re-oxidation of MTZ nitroradical anions. *G. duodenalis* is a microaerophilic parasite that lives in the fairly aerobic (up to 60 μM O_2) environment of the upper intestine, so that a drug potentially effective also in presence of O_2 is more attractive for therapeutic use. We show that NBDHEX binds to gG3PD, when administered both *in vivo* and *in vitro*. NBDHEX is an excellent electrophile that can easily form σ-complexes, as occurs between glutathione and the C-4 position of the drug in the G-site of GST, leading to the inhibition of the detoxifying and anti-apoptotic activity of human GSTP1-1 (Ricci et al., 2005; Federici et al., 2009). Since no gene coding for GST is present in *G. duodenalis* (Morrison et al., 2007), we can *bona fide* exclude that antigiardial activity of NBDHEX may be due to any GST inhibition, although a low level of GSH has been detected (Krauth-Siegel and Leroux, 2012). The formation of σ-complexes likely occurs between NBDHEX and gG3PD, as suggested by in gel fluorescence of gG3PD after NBDHEX treatment, but, unfortunately, our attempts to identify these complexes by mass spectrometry were unsuccessful. Furthermore, we demonstrate the occurrence of NBDHEX covalent adducts at several gG3PD cysteine residues, with the 7-nitro group of the drug being reduced either completely to amine or partially to hydroxylamine. Different NBDHEX modified cysteine residues of His-gG3PD were found in different experimental conditions, i.e. NBDHEX treatment *in E. coli* or *in vitro*, likely as the result of a differential accessibility of the drug to such residues. Our in vitro assays and MS data indicate that the nitro reduced NBDHEX adducts derive from the gG3PD-mediated electrochemical reduction of the drug, possibly via the formation of highly reactive intermediates, such as nitroso or hydroxylamine radicals. It is well known that during the FAD-mG3PD-mediated oxidation of the 2-hydroxy group of g3p, to form DHAP, two electrons are transferred to FAD and then to the quinone pool (Mráček et al., 2013). Similarly, gG3PD activity can generate two electrons able to reduce the nitro group to a hydroxylamine intermediate.

685 Enzymatic reduction has also been proposed as a mechanism for the activation of nitrocompounds, i.e.
686 MTZ, leading to protein covalent adducts formation and cytotoxicity in several microaerophilic
687 protozoan parasites. In particular, the flavoprotein thioredoxin reductase (TRxR) can nitroreduce MTZ,
688 by a process involving its flavin cofactor, thus forming covalent adducts with the drug (Leitsch et al.,
689 2007; Leitsch et al., 2009; Leitsch et al., 2011). Remarkably, although gG3PD displays nitroreductase
690 activity toward NBDHEX, no nitroreduction of MTZ could be observed in our experimental
691 conditions. Nitroimidazoles and other nitrocompounds seem to be metabolized in *G. duodenalis* in a
692 different manner, also in comparison to *E. hystolitica* and *T. vaginalis,* and protein adducts can be
693 formed only with some of these drugs (Leitsch et al., 2012). Indeed, gG3PD was not identified among
694 the proteins forming adducts with MTZ or tinidazole neither in *G. duodenalis* nor in *E. hystolitica*
695 (Leitsch et al., 2012; Leitsch et al., 2007). In particular, these differences in nitrocompound
696 metabolization have been related to the adaptation of *G. duodenalis* to rather high concentrations of O_2
697 in the upper small intestine. We can speculate that gG3PD-mediated NBDHEX nitroreduction may be
698 influenced by a redox-cycling property of 7-nitro benzoxadiazole derivative in presence of O_2. NBF-
699 SPh, a compound related to NBDHEX, has a redox-cycling activity with nitro group cycling to nitroso
700 and hydroxylamine intermediates by reversible electron transfer to O_2 before becoming exhausted and
701 finally reduced to amine (Patridge et al., 2012). During redox-cycling, NBF-SPh rapidly converts O_2 to
702 reactive oxygen species (ROS) at an higher rate than that of the most toxic redox-cycling quinones and
703 the nitrocompound MTZ. Bio-reduction of the redox cycling naphthoquinone menadione by
704 flavoprotein/iron-sulphur-mediated electron transfer in both *G. duodenalis* trophozoites and cysts has
705 been proven to be a potent generator of ROS, even in the presence of low levels of oxygen, thus
706 leading to parasite killing (Paget et al., 2004). Therefore, we propose that the cytotoxicity exerted by
707 NBDHEX against *G. duodenalis* could be due to its nitroreduction and to ROS generation during the
708 reductive process, likely involving gG3PD. Finally, it should be considered that nitroreduction of
709 NBDHEX, at least in part mediated by gG3PD, could subtract electrons from their physiological
710 acceptor(s) thus altering the *G. duodenalis* intracellular redox metabolism.
711 Although further studies are necessary to fully understand the metabolic pathways involving the
712 gG3PD, as well as the mechanisms of the enzyme regulation in *G. duodenalis*, we have clearly proven,
713 for the first time, that a functional FAD-dependent glycerol-3-phospate dehydrogenase works in *G.*
714 *duodenalis* and could be a valuable potential candidate for the design of novel antigiardial drugs having
715 NBDHEX, a representative of novel class of antitumoral molecules, as leading compound.
716

Acknowledgments

We are grateful to Dr. Marialuisa Casella, Istituto Superiore di Sanità, Rome (IT), for the priceless technical assistance with mass spectrometry analyses; to Dr. Pavel Dolezal, Charles University, Prague (CZ), for the gift of the anti-Tom40 antibody; to Dr. Ed Jarroll, Lehman College NY (USA) and Dr. Harry van Keulen, Cleveland State University (USA) for providing us with anti-gPGM antibody. We thanks Prof. Norbert Müller, University of Berne (CH), for the help in drug testing and critical reading of the manuscript and Dr. Donatella Pietraforte, Istituto Superiore di Sanità, Rome (IT), for the fruitful criticism. We also acknowledge Dr. Raffaele Fabrini and Dr. Alessio Bocedi, University of Rome "Tor Vergata" for NDBHEX supply. This study has been partially supported by Istituto Superiore di Sanità (project 14A1/678/2014 and Onco-Technology Program project 13ONC/5).

Bibliography

1. Adam, R.D. (2001). Biology of Giardia lamblia. *Clin. Microbiol. Rev.* 14, 447-475.
2. Albertyn, J., Hohmann, S., Thevelein, J.M., and Prior, B.A. (1994). GPD1, which encodes glycerol-3-phosphate dehydrogenase, is essential for growth under osmotic stress in Saccharomyces cerevisiae, and its expression is regulated by the high-osmolarity glycerol response pathway. *Mol. Cell. Biol.* 14, 4135-4144.
3. Alsmark, U.C., Sicheritz-Ponten, T., Foster, P.G., Hirt, R.P., and Embley, T.M. (2009). Horizontal gene transfer in eukaryotic parasites: a case study of Entamoeba histolytica and Trichomonas vaginalis. *Methods Mol. Biol.* 532, 489-500.
4. Andersson, J.O., Sjögren, A.M., Horner, D.S., Murphy, C.A., Dyal, P.L., Svärd S.G., *et al.* (2007). A genomic survey of the fish parasite Spironucleus salmonicida indicates genomic plasticity among diplomonads and significant lateral gene transfer in eukaryote genome evolution. *BMC Genomics.* 8: 51.
5. Argüello-Garciá, R., Bazán-Tejeda, M.L., and Ortega-Pierres, G. (2009). Encystation commitment in Giardia duodenalis: a long and winding road. *Parasite.* 16, 247-258.
6. Argyrou, A., and Blanchard, J.S. (2004). Flavoprotein disulfide reductases: advances in chemistry and function. *Prog. Nucleic Acid Res. Mol. Biol.* 78, 89-142.
7. Bell, R.M., and Coleman, R.A. (1980). Enzymes of glycerolipid synthesis in eukaryotes. *Annu. Rev. Biochem.* 49, 459-487.
8. Bénéré, E., Inocencio da Luz, R.A., Vermeesch, M., Cos, P., and Maes, L. (2007). A new quantitative in vitro microculture method. *J. Microbiol. Methods.* 71, 101-106.

9. Bragg, P.D., and Hou, C. (1999). Effect of NBD chloride (4-chloro-7-nitrobenzo-2-oxa-1,3-diazole) on the pyridine nucleotide transhydrogenase of Escherichia coli. *Biochim. Biophys. Acta.* 1413, 159-171.
10. Chen, H., Lu, Y., and Jiang, J.G. (2012). Comparative analysis on the key enzymes of the glycerol cycle metabolic pathway in Dunaliella salina under osmotic stresses. *PLoS One.* 7:e37578. doi: 10.1371/journal.pone.0037578
11. Chowdhury, S.K., Raha, S., Tarnopolsky, M.A., and Singh, G. (2007) Increased expression of mitochondrial glycerophosphate dehydrogenase and antioxidant enzymes in prostate cancer cell lines/cancer. *Free Radic. Res.* 41, 1116-1124.
12. Cole, S.T., Eiglmeier, K., Ahmed, S., Honore, N., Elmes, L., Anderson, W.F., *et al.* (1988). Nucleotide sequence and gene-polypeptide relationships of the glpABC operon encoding the anaerobic sn-glycerol-3-phosphate dehydrogenase of Escherichia coli K-12. *J. Bacteriol.* 170, 2448-2456
13. Dagley, M.J., Dolezal, P., Likic, V.A., Smid, O., Purcell, A.W., Buchanan, S.K., *et al.* (2009). The protein import channel in the outer mitosomal membrane of Giardia intestinalis. *Mol. Biol. Evol.* 26, 1941-1947.
14. Dinkel H, Van Roey K, Michael S, Davey NE, Weatheritt RJ, Born D, *et al.* (2014). The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Res.* 42 (Database issue):D259-66. doi: 10.1093/nar/gkt1047.
15. Dolezal, P., Smíd, O., Rada, P., Zubácová, Z., Bursać, D., Suták, R., *et al.* (2005). Giardia mitosomes and trichomonad hydrogenosomes share a common mode of protein targeting. *Proc. Natl. Acad. Sci. USA.* 102, 10924-10929.
16. Dolgikh, V.V., Senderskiy, I.V., Pavlova, O.A., Naumov, A.M., and Beznoussenko, G.V. (2011). Immunolocalization of an alternative respiratory chain in Antonospora (Paranosema) locustae spores: mitosomes retain their role in microsporidial energy metabolism. *Eukaryot. Cell.* 10, 588-593.
17. Federici, L., Lo Sterzo, C., Pezzola, S., Di Matteo, A., Scaloni, F., Federici, G., *et al.* (2009). Structural basis for the binding of the anticancer compound 6-(7-nitro-2,1,3-benzoxadiazol-4-ylthio)hexanol to human glutathione s-transferases. *Cancer Res.* 69, 8025-8034.
18. Gardino, A.K., and Yaffe, M.B. (2011). 14-3-3 proteins as signaling integration points for cell cycle control and apoptosis. *Semin. Cell. Dev. Biol.* 22, 688-695.

780 19. Gault, M.J., Gillin, F.D., and Zenian, A.J. (1987). Giardia lamblia: stimulation of growth by human
781 intestinal mucus and epithelial cells in serumfree medium. *Exp. Parasitol.* 64, 29-37.
782 20. Gillin, F.D., Gault, M.J., Hofmann, A.F., Gurantz, D., and Sauch, J.F. (1986). Biliary lipids support
783 serum-free growth of Giardia lamblia. *Infect. Immun.* 53, 641-645.
784 21. Großhennig, S., Schmidl, S.R., Schmeisky, G., Busse, J., and Stülke, J. (2013). Implication of
785 glycerol and phospholipid transporters in Mycoplasma pneumoniae growth and virulence. *Infect.*
786 *Immun.* 81, 896-904.
787 22. Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis. *Nucl.*
788 *Acids. Symp.* Ser. 41, 95-98
789 23. Halliez, M.C, and Buret, A.G. (2013). Extra-intestinal and long term consequences of Giardia
790 duodenalis infections. *World J. Gastroenterol.* 19, 8974-8985.
791 24. Han, J., and Collins, L.J. (2012). Reconstruction of Sugar Metabolic Pathways of Giardia lamblia.
792 *Int. J. Proteomics.* 2012:980829. doi: 10.1155/2012/980829.
793 25. Hounkong, K., Sawangjaroen, N., and Phongpaichit, S. (2011). A colorimetric method for the
794 evaluation of anti-giardial drugs in vitro. *Exp. Parasitol.* 127, 600-603.
795 26. Husain, A., Sato, D., Jeelani, G., Soga, T., and Nozaki, T. (2012). Dramatic increase in glycerol
796 biosynthesis upon oxidative stress in the anaerobic protozoan parasite Entamoeba histolytica. *PLoS*
797 *Negl. Trop. Dis.* 6:e1831. doi: 10.1371/journal.pntd.0001831.
798 27. Janssen, M.J., van Voorst, F., Ploeger, G.E., Larsen, P.M., Larsen, M.R., de Kroon, A.I., *et al.*
799 (2002). Photolabeling identifies an interaction between phosphatidylcholine and glycerol-3-
800 phosphate dehydrogenase (Gut2p) in yeast mitochondria. *Biochemistry* 41, 5702-5711.
801 28. Jarroll, E.L., Muller, P.J., and Meyer, E.A., and Morse, S.A. (1981). Lipid and carbohydrate
802 metabolism of Giardia lamblia. *Mol. Biochem. Parasitol.* 2, 187-196.
803 29. Jedelský, P.L., Dolezal, P., Rada, P., Pyrih, J., Smíd, O., Hrdý, I., *et al.* (2011). The minimal
804 proteome in the reduced mitochondrion of the parasitic protest. Giardia intestinalis. *PLoS One.*
805 6:e17285. doi: 10.1371/journal.pone.0017285.
806 30. Kistler, W.S., and Lin, E.C. (1972). Purification and properties of the flavine-stimulated anaerobic
807 L- -glycerophosphate dehydrogenase of Escherichia coli. *J. Bacteriol.* 112, 539-547.
808 31. Kleppe, R., Martinez, A., Døskeland, S.O., and Haavik, J. (2011). The 14-3-3 proteins in regulation
809 of cellular metabolism. *Semin. Cell. Dev. Biol.* 22, 713-719.
810 32. Krauth-Siegel, R.L., and Leroux, A.E. (2012). Low-molecular-mass antioxidants in parasites.
811 *Antioxid. Redox Signal.* 17, 583-607.

812 33. Lalle, M., Salzano, A.M., Crescenzi, M., and Pozio, E. (2006). The Giardia duodenalis 14-3-3
813 protein is post-translationally modified by phosphorylation and polyglycylation of the C-terminal
814 tail. *J. Biol. Chem.* 281, 5137-5148.
815 34. Lalle, M. (2010). Giardiasis in the post genomic era: treatment, drug resistance and novel
816 therapeutic perspectives. *Infect. Disord. Drug Targets.* 10:283–294.
817 35. Lalle, M., Camerini, S., Cecchetti, S., Sayadi, A., Crescenzi, M., and Pozio, E. (2012). Interaction
818 network of the 14-3-3 protein in the ancient protozoan parasite Giardia duodenalis. *J. Proteome Res.*
819 11, 2666-2683.
820 36. Leitsch, D., Kolarich, D., Wilson, I.B., Altmann, F., and Duchêne, M. (2007). Nitroimidazole action
821 in Entamoeba histolytica: a central role for thioredoxin reductase. *PLoS Biol.* 5:e211.
822 37. Leitsch, D., Kolarich, D., Binder, M., Stadlmann, J., Altmann, F., and Duchêne, M. (2009).
823 Trichomonas vaginalis: metronidazole and other nitroimidazole drugs are reduced by the flavin
824 enzyme thioredoxin reductase and disrupt the cellular redox system. Implications for nitroimidazole
825 toxicity and resistance. *Mol. Microbiol.* 72, 518-536.
826 38. Leitsch, D., Burgess, A.G., Dunn, L.A., Krauer, K.G., Tan, K., Duchêne, M., *et al.* (2011).
827 Pyruvate:ferredoxin oxidoreductase and thioredoxin reductase are involved in 5-nitroimidazole
828 activation while flavin metabolism is linked to 5-nitroimidazole resistance in Giardia lamblia. *J.*
829 *Antimicrob. Chemother.* 66, 1756-1765.
830 39. Leitsch, D., Schlosser, S., Burgess, A., and Duchêne, M. (2012). Nitroimidazole drugs vary in their
831 mode of action in the human parasite Giardia lamblia. *Int. J. Parasitol. Drugs Drug Resist.* 2:166-
832 70.
833 40. Lopez, A.B., Sener, K., Jarroll, E.L., and van Keulen, H. (2003). Transcription regulation is
834 demonstrated for five key enzymes in Giardia intestinalis cyst wall polysaccharide biosynthesis.
835 *Mol. Biochem. Parasitol.* 128, 51-57.
836 41. Miki, K., and Wilson, T.H. (1978). Proton translocation associated with anaerobic
837 transhydrogenation from glycerol 3-phosphate to fumarate in Escherichia coli. *Biochem. Biophys.*
838 *Res. Commun.* 83, 1570-1575.
839 42. Morrison, H.G., McArthur, A.G., Gillin, F.D., Aley, S.B., Adam, R.D., Olsen, G.J., *et al.* (2007).
840 Genomic minimalism in the early diverging intestinal parasite Giardia lamblia. *Science.* 317, 1921-
841 1926.
842 43. Mráček, T., Drahota, Z., and Houštěk, J. (2013). The function and the role of the mitochondrial
843 glycerol-3-phosphate dehydrogenase in mammalian tissues. *Biochim. Biophys. Acta.* 1827, 401-410.

844 44. Müller, J., and Hemphill, A. (2013). New approaches for the identification of drug targets in
845 protozoan parasites. *Int. Rev. Cell. Mol. Biol.* 301, 359-401.
846 45. Paget, T.A., Macechko, P.T., and Jarroll, E.L. (1998). Metabolic changes in Giardia intestinalis
847 during differentiation. *J. Parasitol.* 84, 222-226.
848 46. Paget, T., Maroulis, S., Mitchell, A., Edwards, M.R., Jarroll, E.L., and Lloyd, D. (2004). Menadione
849 kills trophozoites and cysts of Giardia intestinalis. *Microbiology.* 150, 1231-1236.
850 47. Patridge, E.V., Eriksson, E.S., Penketh, P.G., Baumann, R.P., Zhu, R., Shyam, K., *et al.* (2012). 7-
851 Nitro-4-(phenylthio)benzofurazan is a potent generator of superoxide and hydrogen peroxide. *Arch.*
852 *Toxicol.* 86, 1613-1625.
853 48. Pearson, W.R., and Lipman, D.J. (1988). Improved tools for biological sequence comparison. *Proc.*
854 *Natl. Acad. Sci. USA.* 85, 2444-2448.
855 49. Pellizzari Tregno, F, Sau, A, Pezzola, S, Geroni, C, Lapenta, C, Spada, M, *et al.* (2009). In vitro and
856 in vivo efficacy of 6-(7-nitro-2,1,3-benzoxadiazol-4-ylthio)hexanol (NBDHEX) on human
857 melanoma. *Eur. J. Cancer.* 45, 2606-2617.
858 50. Powell, S., Szklarczyk, D., Trachana, K., Roth, A., Kuhn, M., Müller, J., *et al.* (2012). eggNOG
859 v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic. Acids*
860 *Res.* 40 (Database issue):D284-D289.
861 51. Pyrih, J., Harant, K., Martincová, E., Sutak, R., Lesuisse, E., Hrdý, I., *et al.* (2014). Giardia
862 intestinalis incorporates heme into cytosolic cytochrome b_5. *Eukaryot. Cell.* 13, 231-239.
863 52. Ricci, G., De Maria, F., Antonini, G., Turella, P., Bullo, A., Stella, L., *et al.* (2005). 7-Nitro-2,1,3-
864 benzoxadiazole derivatives, a new class of suicide, inhibitors for glutathione S-transferases. Mec
865 53. Ryan, U., and Cacciò, S.M. (2013). Zoonotic potential of Giardia. *Int. J. Parasitol.* 43, 943-956.
866 54. Schofield, P.J., Edwards, M.R., and Kranz, P. (1991). "Glucose metabolism in Giardia intestinalis."
867 *Mol. Biochem. Parasitol.* 45, 39-47
868 55. Schupp, D.G., Januschka, M.M., Sherlock, L.A., Stibbs, H.H., Meyer, E.A., Bemrick, W.J., (1988)
869 *et al.* Production of viable Giardia cysts in vitro: determination by fluorogenic dye staining,
870 excystation, and animal infectivity in the mouse and Mongolian gerbil. *Gastroenterology.* 95, 1-10.
871 56. Singh, G. (2014). Mitochondrial FAD-linked Glycerol-3-phosphate Dehydrogenase: A Target for
872 Cancer Therapeutics. *Pharmaceuticals (Basel).* 7, 192-206.
873 57. Sousa M.C., Gonçalves C.A., Baires V.A., Poiares-da-Silva J. (2001). Adherence of Giardia lamblia
874 trophozoites to Int-407 human intestinal cells. *Clin. Diagn. Lab. Immunol.* 8, 258-265.

875 58. Stevens, T.L., Gibson, G.R., Adam. R., Maier. J., Allison-Ennis. M., and Das. S. (1997). Uptake and
876 cellular localization of exogenous lipids by Giardia lamblia, a primitive eukaryote. *Exp. Parasitol.*
877 86, 133-143.
878 59. Suescún-Bolívar, L.P., and Thomé, P.E. (2015). Osmosensing and osmoregulation in unicellular
879 eukaryotes. *World J. Microbiol. Biotechnol.* 31, 435-443.
880 60. Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA5: Molecular
881 Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum
882 Parsimony Methods. *Mol. Biol. Evol.* 28, 2731-2739.
883 61. Tentori, L., Dorio, A.S., Mazzon, E., Muzi, A., Sau, A., Cuzzocrea, S., *et al.* (2011). The glutathione
884 transferase inhibitor 6-(7-nitro-2,1,3-benzoxadiazol-4-ylthio)hexanol (NBDHEX) increases
885 temozolomide efficacy against malignant melanoma. *Eur. J. Cancer.* 47, 1219-1230.
886 62. Tovar, J., León-Avila, G., Sánchez, L.B., Sutak, R., Tachezy, J., van der Giezen, M., *et al.* (2003).
887 Mitochondrial remnant organelles of Giardia function in iron-sulphur protein maturation. *Nature.*
888 426, 172-176.
889 63. Turella, P., Cerella, C., Filomeni, G., Bullo, A., De Maria, F., Ghibelli, L., *et al.* (2005).
890 Proapoptotic activity of new glutathione S-transferase inhibitors. *Cancer Res.* 65, 3751-3761.
891 64. Unden, G., and Bongaerts, J. (1997). Alternative respiratory pathways of Escherichia coli:
892 energetics and transcriptional regulation in response to electron acceptors. *Biochim. Biophys. Acta.*
893 1320, 217-234.
894 65. Varga, M.E., and Weiner, J.H. (1995). Physiological role of GlpB of anaerobic glycerol-3-phosphate
895 dehydrogenase of Escherichia coli. *Biochem. Cell. Biol.* 73, 147-153.
896 66. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., Barton, G.J. (2009). Jalview Version 2--
897 a multiple sequence alignment editor and analysis workbench. Bioinformatics. 25, 1189-1191.
898 67. Watkins, R.R., and Eckmann, L. (2014). Treatment of giardiasis: current status and future directions.
899 *Curr. Infect. Dis. Rep.* 16, 396. doi: 10.1007/s11908-014-0396-y.
900 68. Yang, W., Cao, Y., Sun, X., Huang, F., He, Q., Qiao, D., *et al.* (2007). Isolation of a FAD-GPDH
901 gene encoding a mitochondrial FAD-dependent glycerol-3-phosphate dehydrogenase from
902 Dunaliella salina. *J. Basic Microbiol.* 47, 266-274.
903 69. Yeh, J.I., Chinte, U., and Du, S. (2008). Structure of glycerol-3-phosphate dehydrogenase, an
904 essential monotopic membrane enzyme involved in respiration and metabolism. *Proc. Natl. Acad.*
905 *Sci. USA.* 105, 3280-3285.

906 70. Yichoy, M., Duarte, T.T., De Chatterjee, A., Mendez, T.L., Aguilera, K.Y., Roy, D., *et al.* (2011).
907 Lipid metabolism in Giardia: a post-genomic perspective. *Parasitology.* 138, 267-278.

908 **Figure legends**

909

910 **Figure 1. Expression and localization of the gG3PD during the differentiation stages of *G.***
911 ***duodenalis*.** A) Western blot from three independent analysis of Triton lysates (20 µg) from *G.*
912 *duodenalis* WB-C6 trophozoites (T) and parasites harvested at 6, 12 and 24h after encystation
913 induction. Immunoblotting was performed with: anti-gG3PD polyclonal serum (α-gG3PD); anti-
914 phosphoacetylglucosamine mutase (α-PGM) (Lopez et al., 2003), to follow the progression of
915 encystation; the anti-α-tubulin (α-α-TUB), as loading control. Molecular size markers (kDa) are
916 reported on the left. The analysis is representative of three independent experiments. B) CLSM
917 observations of fixed and permeabilized *G. duodenalis* WB-C6 parasites at different stages: trophozoite
918 (panel a, Throph.), encysting parasite after 12h of encystation (panel b and c, Encyst.) and cyst (panel
919 d, Cyst) stained with the mouse polyclonal serum α-gG3PD (red) and the α-g14-3-3 rabbit polyclonal
920 (green). Cyst wall and encystation specific vesicles (ESVs) were stained with Cy3-conjugated α-CWP
921 mAb (grey). Nuclei (N) were stained with DAPI (blue). Displayed micrographs correspond to a single
922 z-stack: a and b, ventral stacks; c and d, central stacks encompassing the nuclei. T, transmission light
923 acquisition. Scale bars, 5µm. Arrows indicate the ventrolateral flanges (VLF). A magnification (zoom)
924 of the indicated area in the merged image is shown. C) CLSM observation as in B. Parasites were
925 stained with α-gG3PD (red) and α-Tom40 antiserum (green) (Dagley et al., 2009). Nuclei (N) were
926 stained with DAPI (blue). A magnification (zoom) of the indicated area in the merged image is shown.
927 Mitosomes are indicated (m). Images in B) and C) are representative of >50 fields analyzed in two
928 independent experiments.

929

930 **Figure 2. Expression of the recombinant FLAG-HA-gG3PD in *G. duodenalis* and its interaction**
931 **with g14-3-3.** A) Western blot analysis of Triton lysates from trophozoites (T) or parasites after 12h of
932 encystation (E) obtained from FLAG-HA-gG3PD transfected line or from the control WBC6 strain. 20
933 µg of protein extracts were loaded in each lane and separated on a 4-12% SDS-PAGE, then transferred
934 on nitrocellulose membrane and finally probed with the indicated antibodies. Equal protein loading was
935 monitored by anti-TUB Ab, whereas encystation induction was confirmed by anti-gPGM Ab.
936 Molecular size markers (kDa) are reported in the left. B) Subcellular distribution of gG3PD. 30 µg of
937 fractionated protein lysate, soluble (S) or octyl β-D-glucopyranoside solubilized membrane proteins
938 (M), from trophozoites (T) or 12h encysting parasites (E) of FLAG-HA-gG3PD transfected parasites or
939 control WB-C6, were treated as described in (A) and probed with the indicated antibodies. Molecular

size markers are on the left. C) Co-immunoprecipitation assay of endogenous g14-3-3 with FLAG-HA-gG3PD. An aliquot (1:5) of the FLAG peptide-eluted material from WB-C6 or FLAG-HA-gG3PD transgenic line, deriving from trophozoites (T) or encysting parasite (E), was separated on 4-12% SDS-PAGE and immunoblotted with anti-HA (α-HA) or anti-g14-3-3 polyclonal serum (α-g14-3-3). Molecular size markers are on the left. Immunoblots of panel A), B) and C) are representative of three independent experiments.

Figure 3. Expression and intracellular localization of the FLAG-HA-tagged gG3PD deletion mutants in *G. duodenalis* parasites. A) Representative western blot analysis of soluble protein lysate (20 µg) from control WB-C6 and transgenic trophozoites (T) or parasites after 12h of encystation (E), expressing the full-length FLAG-HA-G3PD, or the FLAG-HA-gG3PD_N or _C. Lysates were separated on 4-12% SDS-PAGE and immunoblotted with the indicated antibodies. Arrows on the right indicate the molecular size of the corresponding proteins. Molecular size markers are on the left. The vertical line in the panel corresponding to the α-HA blotting indicates two different times of exposure between the samples of the same gel. The immunoblots are representative of three independent experiments. B) CLSM observations of fixed and permeabilized transgenic *G. duodenalis* parasite expressing the FLAG-HA-gG3PD_N or the FLAG-HA-gG3PD_C at different stages: trophozoite (panels a and d, Throph.), encysting parasites after 12h of encystation (panels b and e, Encyst.) and cysts (panels c and f, Cyst). Parasites were stained with mouse α-HA mAb (green) and rabbit polyclonal α-g14-3-3 (red). Cyst wall and encystation specific vesicles (ESVs) were stained by Cy3-conjugated α-CWP mAb (grey). Nuclei were DAPI-stained (blue). Displayed micrographs correspond to a single z-stack. T, transmission light acquisition. Scale bars are reported. Arrows indicate the ventral plasma membrane and ventrolateral flanges (VLF). Images are representative of >50 fields analyzed in two independent experiments.

Figure 4. Evaluation of the gG3PD enzymatic activity. A) Spectrofotometric analysis of purified HIS-gG3PD. The UV-visible spectrum of HIS-gG3PD (10 mg/ml) in 67 mM of potassium phosphate buffer, pH 7.5, was recorded at 25°C. The insert shows a magnification of the HIS-gG3PD spectrum (solid line) in comparison with the spectrum of authentic FAD (dotted line) recorded in the same buffer. Peak maxima are reported. Spectra are representative of three independent experiments. B) Assessment of HIS-gG3PD dimerization *in vitro*. Purified recombinant proteins (3 µmol) were separated on 3–12% Blue Native-PAGE and silver-stained or transferred on polyvinylidene difluoride

(PVDF) membrane and probed with anti-HIS mAb. Native size markers (kDa) are indicated on the left. Asterisks indicate HIS-gG3PD monomer (*) or dimer (**). Empty dots indicate HIS-gG3PD_C monomer (°) or dimer (°°). The arrow indicates HIS-gG3PD_N monomer (<). Native-PAGE and immunoblot are representative of three independent experiments. C) The enzymatic activities of both the purified HIS-gG3PD and the deletion mutant HIS-gG3PD_N were measured *in vitro*, by MTT assay, in the presence (+FAD) or absence (no FAD) of 10µM FAD. The G3PD activity (mean ± SD) from three independent experiments is expressed as nmol of reduced MTT per min per mg of recombinant protein. Statistical analyses were performed using unpaired t-test between the full length and the deletion mutant: * P< 0.05 and ** P< 0.01. D) The FAD-glycerol-3-phosphate dehydrogenase activity was measured in protein extract (100 µg) from *G. duodenalis* trophozoites (Troph.) or in parasite after 6h or 12h from encystation induction (Encyst.). The relative enzymatic activity (mean ± SD) from three independent experiments is expressed as the percentage change respect to the value measured in trophozoites. Statistical analyses were performed using unpaired t-test (Troph vs Encyst 6h (** P<0.01) and Troph vs Encyst 12h (ns)), and one-way ANOVA (°° P<0.001). E) Western blot analysis of protein extracts (20 µg) used to assay the gG3PD enzymatic activity (as described in D), extracts were separated on 4-12% SDS-PAGE and immunoblotted with the indicated antibodies. Molecular size markers are indicated on the left. Immunoblot is representative of three independent experiments. F) The intracellular glycerol content was measured in supernatant from *G. duodenalis* trophozoites (Troph) and parasites after 6h or 12h from encystation induction (Encyst.). The relative glycerol amount of three independent experiments (mean ± SD) is expressed as the percentage change respect to the amount detected in trophozoites. Statistical analyses were performed using unpaired t-test between Troph and Encyst 6h (*** P<0.001) and Troph vs Encyst 12h (** P<0.01). One-way ANOVA confirmed that differences among groups were statistically significant (°° P<0.001).

Figure 5. Evaluation of NBDHEX effects on *G. duodenalis* growth and gG3PD enzymatic activity. A) Survival of *G. duodenalis* WBC6 trophozoites was determined by methylene blue colorimetric assay after 48h of treatment with different concentrations, ranging from 0.05 to 10 µM, of NBDHEX (empty dots) or metronidazole (MTZ, empty squares) in microaerophilic growth conditions. Data (mean percentage ± SD) represent three independent experiments, each done in triplicate. The structure of the NBDHEX compound is reported in the insert. B) The FAD-glycerol-3-phosphate dehydrogenase activity was measured in protein extract (100 µg) from *G. duodenalis* trophozoites treated for the indicated times with 50 µM NBDHEX or for 6h with ethanol (Control). The enzymatic activity of three

1004 independent experiments (mean ± SD) is expressed as the percentage change respect to the control.
1005 Unpaired t-test was performed between control and each time point (* P<0.05 and ** P<0.01). One-
1006 way ANOVA indicated statistically significant differences among all stages (°°° P<0.0001). C) 20 µg
1007 of protein extracts derived from trophozoites treated as described in (B), were separated on 4-12%
1008 SDS-PAGE and immunoblotted. A representative western blot analysis is shown and the antibodies
1009 indicated. Molecular size markers are reported in the left. Table in the bottom reports the densitometric
1010 analysis of three independent experiments (mean ± SD). Statistical analyses using t-test were not
1011 significant.
1012
1013 **Figure 6. Interaction of NBDHEX with gG3PD.** A) CLSM analysis of fixed and permeabilized *G.*
1014 *duodenalis* WBC6 trophozoites after 2h incubation with 50µM NBDHEX. NBDHEX (green) was
1015 directly visualized using the laser light at 488 nm ex. Parasites were stained with rabbit polyclonal anti-
1016 g14-3-3 and AlexaFluor 647-conjugated anti-rabbit (red). Nuclei were stained with DAPI (blue).
1017 Displayed micrographs correspond to a single z-stack. T, transmission light acquisition. Scale bar is
1018 reported. Images are representative of >50 fields analyzed in two independent experiments. B)
1019 Histograms represent the FAD-glycerol-3-phosphate dehydrogenase activity (mean percentage ± SD)
1020 of recombinant HIS-gG3PD (16 pmol) purified from HIS-gG3PD-overproducing *E. coli* after 2h
1021 incubation with 50µM NBDHEX (NBDHEX) or ethanol (Control). The relative change in the
1022 enzymatic activity is given as percentage in relation to the control. Statistical analyses were performed
1023 using unpaired t-test (**** P<0.0001) on three independent experiments. C) SDS-PAGE (4-12%) of 2
1024 µg of purified NBDHEX-treated (NBDHEX) or ethanol-treated (Control) HIS-gG3PD, as described in
1025 (B), under reducing (+DTT) or not reducing condition. On the left, gel stained with Coomassie blue; on
1026 the right, the same gel photographed under UV light prior to staining. Molecular size markers are
1027 indicated on the left. D) MS/MS spectra matching the gG3PD peptides (residues 539-551, left panel,
1028 and residues 859-872, right panel) carrying a mass increase of 265 Da on the cysteine residues C545
1029 and C865, respectively. The mass shift is compatible with an NBDHEX-derived adduct in which the
1030 nitro group was completely reduced to an amine. Data shown in C) and D) are representative of three
1031 independent experiments.
1032
1033 **Figure 7. NBDHEX is substrate of HIS-gG3PD.** A) Spectrophotometric analysis of the reaction
1034 between NBDHEX and gG3PD. The UV-visible spectrum of NBDHEX (100µM) in 67 mM potassium
1035 phosphate buffer (pH 7.5) was recorded at 25°C either in the presence of 17 mM g3p (red line, +g3p)

or 80 nM HIS-gG3PD (dashed black line, +HIS-gG3PD). The NBDHEX (100µM) UV-visible spectrum was also recorded at different time points (0, 10, 20, 40 and 80 min) in the presence of both 17 mM g3p and 80 nM HIS-gG3PD (+HIS-gG3PD+g3p). Arrows indicate the wavelength (nm) of the maximal absorption. B) Fluorescence spectra of 100 µM NBDHEX (excitation at 430 nm) after 80 min of incubation with 80 nM HIS-gG3PD, either in the presence (solid line) or absence (dashed line) of 17 mM g3p. Fluorimetric analyses were recorded at 25°C in 67 mM potassium phosphate buffer (pH 7.5). NBDHEX fluorescence emission has a maximum peak wavelength at 525 nm. Fluorescence intensity (y axis) is arbitrary. The change in color of the two reactions is shown in the insert. C) The UV-visible spectrum of NBDHEX (100µM) in 67 mM potassium phosphate buffer (pH 7.5) was recorded at 25°C before (solid line) and after (dashed line) treatment with sodium dithionite ($Na_2S_2O_4$). The wavelengths (nm) of the maximal absorption are indicated with arrows. D) Fluorescence spectra of 100 µM NBDHEX (excitation at 430 nm) before (solid line) and after $Na_2S_2O_4$ (dashed line). The slight change in color of the two conditions is shown in the insert. Spectra and images from A) to D) are representative of three independent experiments. E) SDS-PAGE (4-12%) of 2 µg of HIS-gG3PD after 80 min of incubation at 25°C with NBDHEX (+NBDHEX) in the presence or not of g3p, or with ethanol (no NBDHEX). Gel was photographed under UV light (UV, right panel) before staining with Coomassie blue (Coomassie, left panel). Molecular size markers are reported in the left. Images are representative of three independent experiments. F) Comparison of the UV-visible spectra of NBDHEX (green and orange line) or MTZ (solid and dashed line) incubated with HIS-gG3PD and g3p (G3PD) at time 0 min and after 80 min of incubation at 25°C (T0 and T80, respectively). Arrows indicate the wavelengths (nm) of the maximal absorption. Spectra are representative of three independent experiments.

1059 **Tables**

1060

1061 **Table 1**. Primer list

Target	Forward primer[a]	Reverse primer[b]
Full length gG3PD	**gG3PDforw** 5′-GGATCCACCACCGCTCACCCTTTACC-3′	**gG3PDrev** 5′-GCGGCCGCTCACTTCATAGCCGGAATGTTC-3′
N-terminal gG3PD	**gG3PDforw**	**NTrev** 5′-GCGGCCGCTCAGTATGCCTTGTCCATCGGGGTG-3′
C-terminal gG3PD	**CTforw** 5′-GGATCCACCCCGATGGACAAGGCATAC-3′	**gG3PDrev**

1062 a The BamHI site is underlined.

1063 b The NotI site is underlined.

Figure 1.TIF

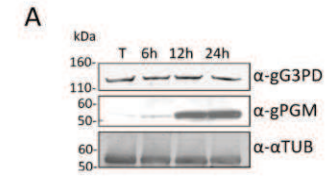


Figure 1

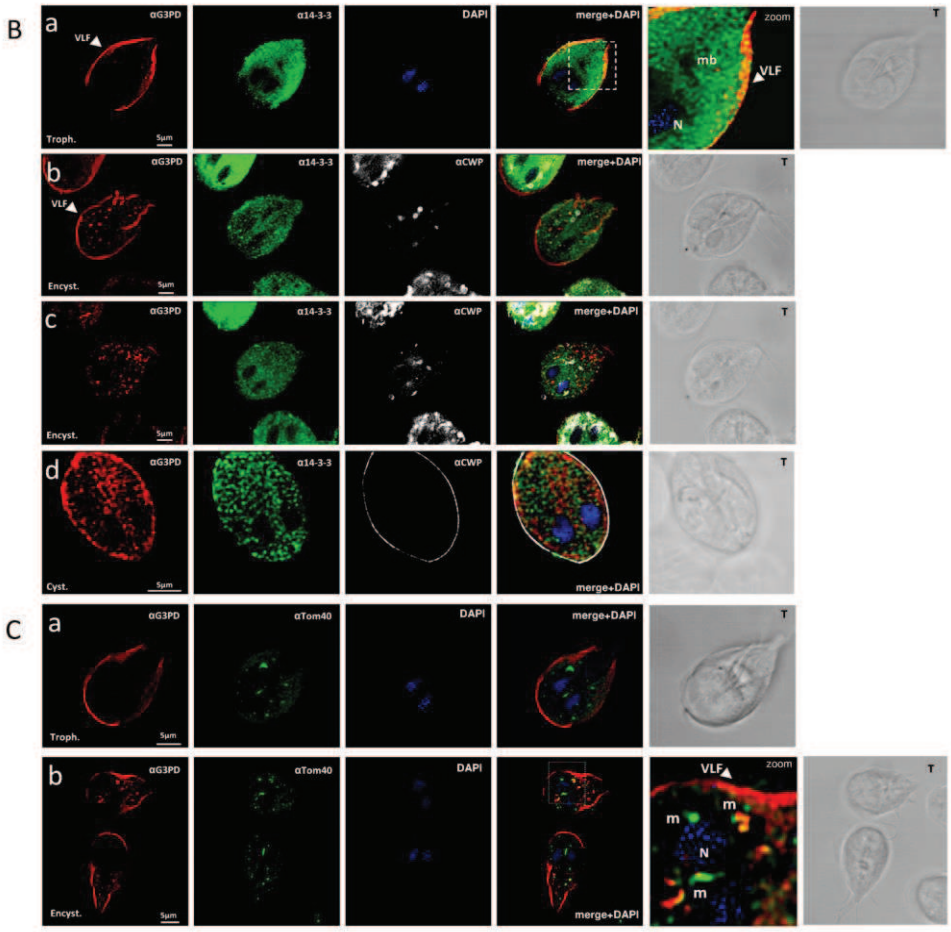


Figure 2.TIF

Figure 2

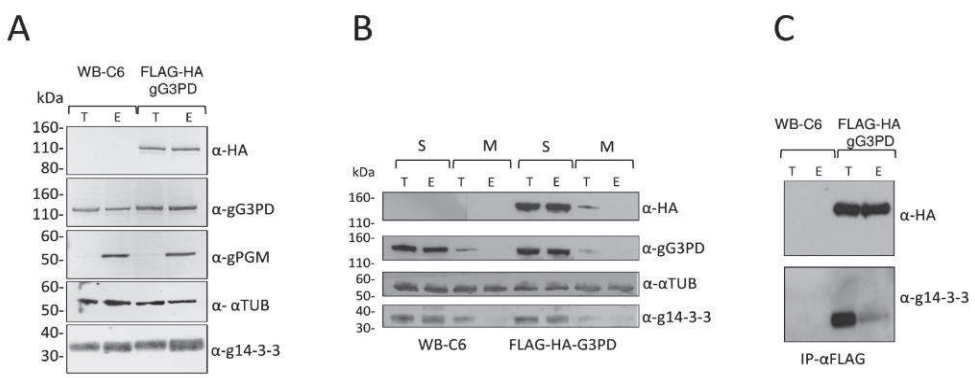


Figure 3.TIF

Figure 3

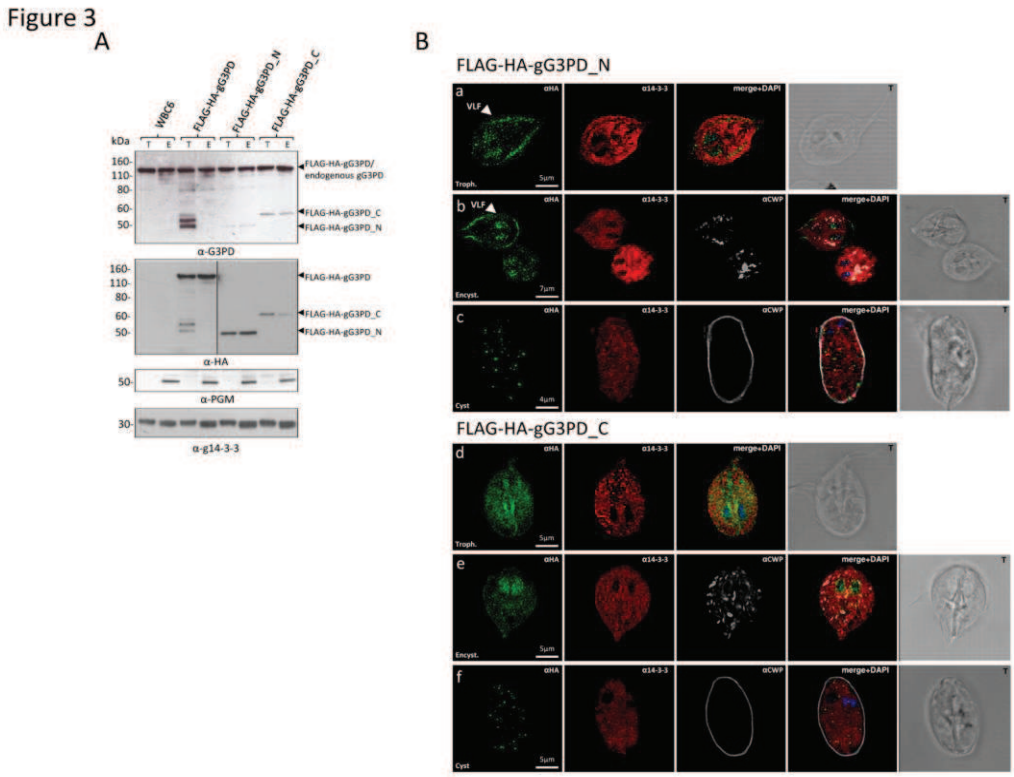


Figure 4.TIF

Figure 4

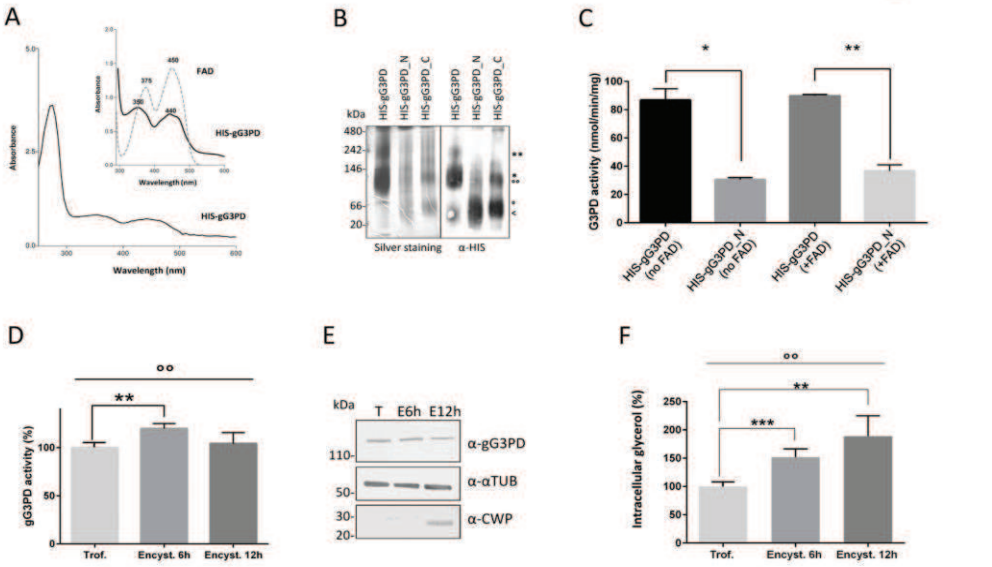


Figure 5.TIF

Figure 5

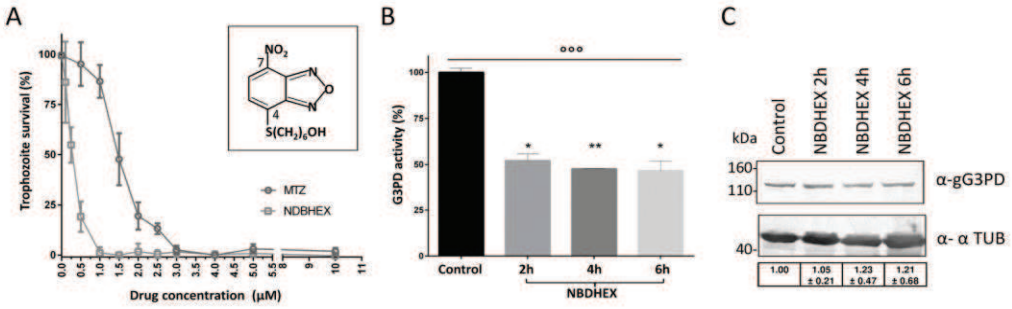


Figure 6.TIF

Figure 6

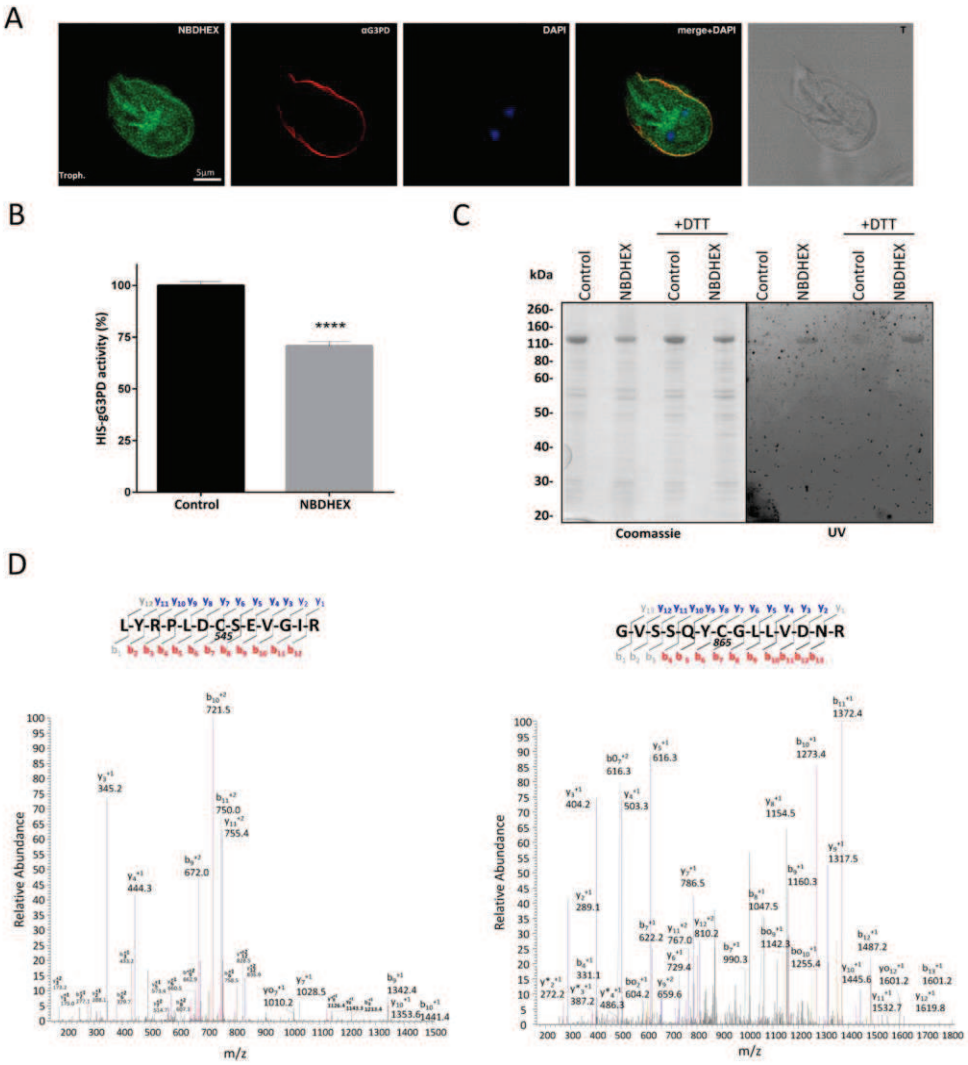


Figure 7.TIF

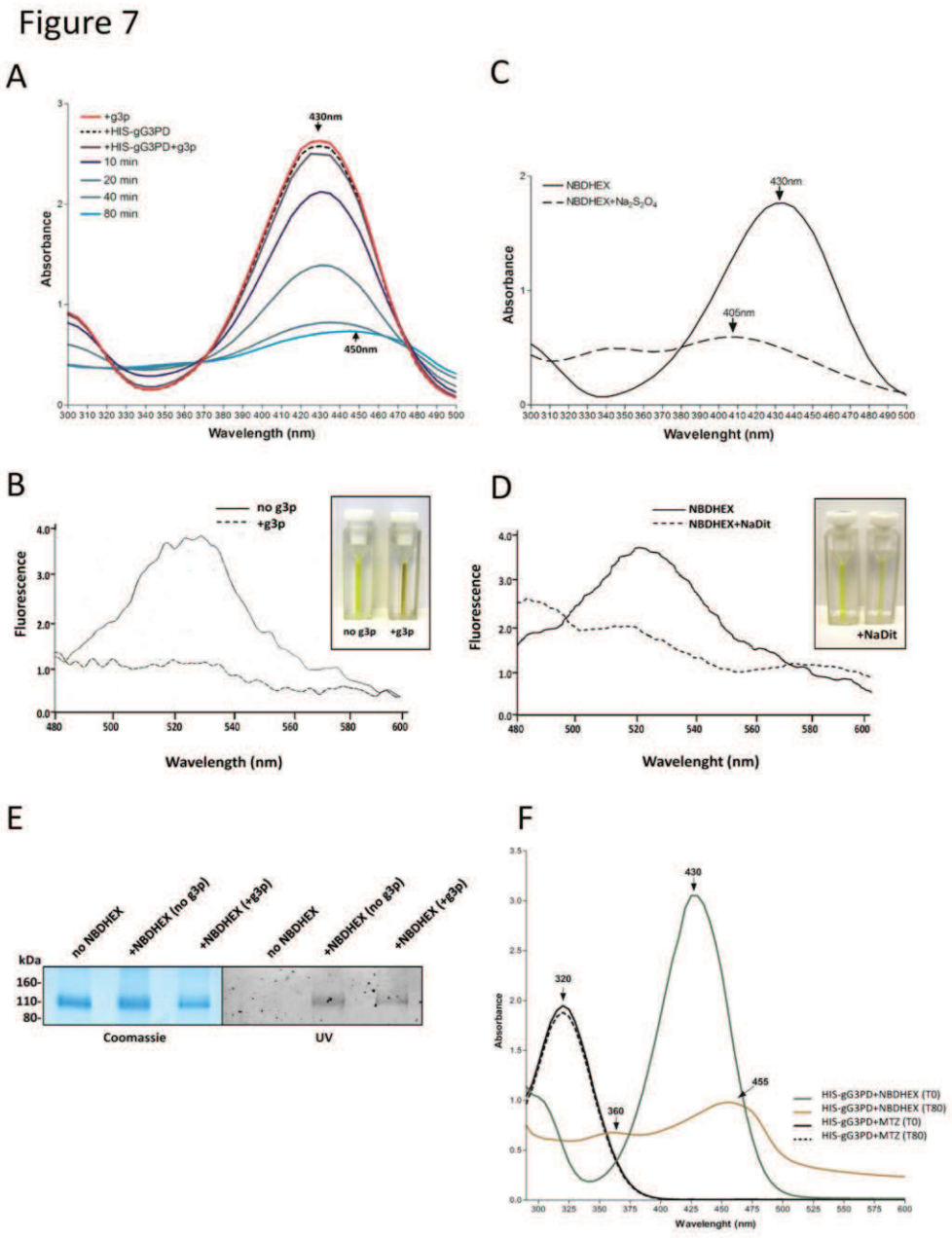
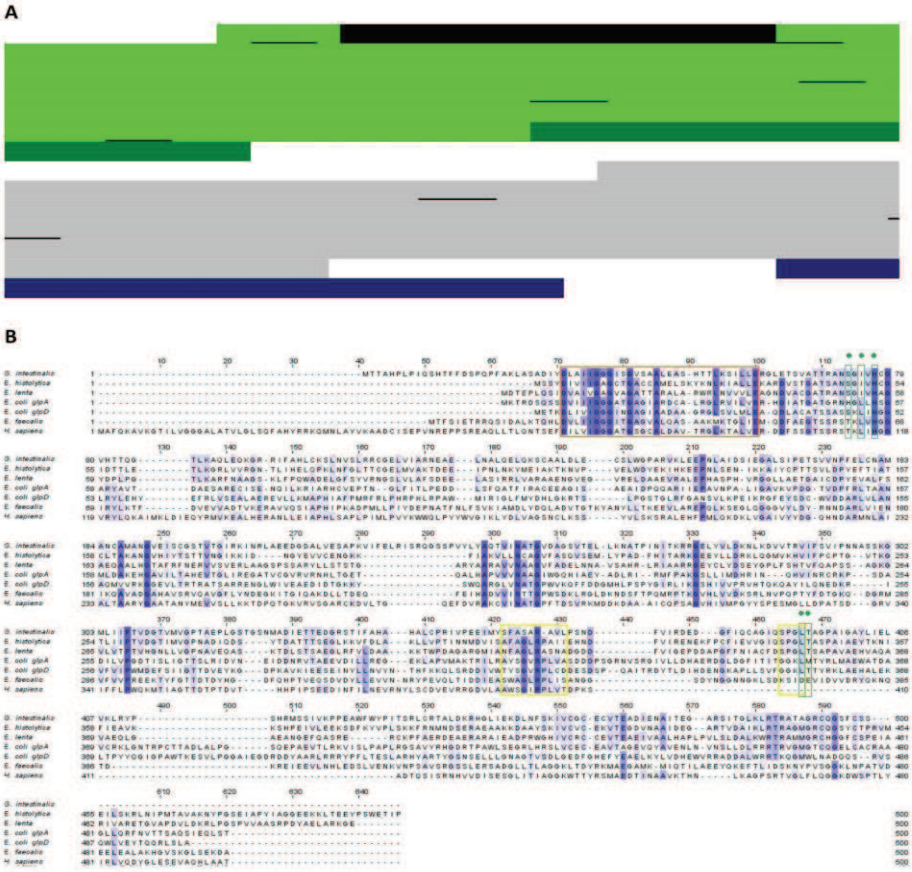


Figure 7
A
+g3p
+HIS-gG3PD
+HIS-gG3PD+g3p
10 min
20 min
40 min
80 min
430nm
450nm
Absorbance
Wavelength (nm)
C
NBDHEX
NBDHEX+Na2S2O4
430nm
405nm
Absorbance
Wavelenght (nm)
B
no g3p
+g3p
Fluorescence
Wavelength (nm)
D
NBDHEX
NBDHEX+NaDit
+NaDit
Fluorescence
Wavelenght (nm)
E
no NBDHEX
+NBDHEX (no g3p)
+NBDHEX (+g3p)
kDa
160-
110-
80-
Coomassie
UV
F
430
320
360
455
Absorbance
Wavelenght (nm)
HIS-gG3PD+NBDHEX (T0)
HIS-gG3PD+NBDHEX (T80)
HIS-gG3PD+MTZ (T0)
HIS-gG3PD+MTZ (T80)

Supplemental Information

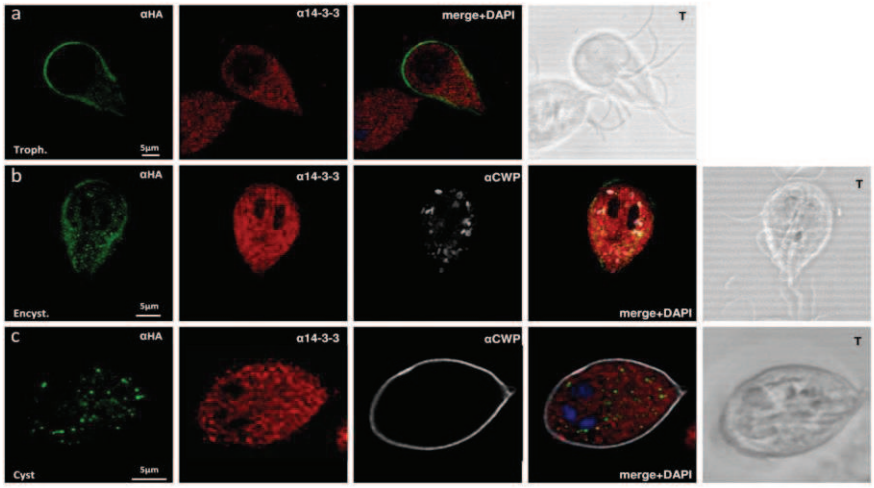
Supplemental Figure S1



Supplemental Figure S1. A) Domain organization of the gG3PD protein. Light and dark green boxes, correspond to the multidomain TIGR03377 of subunit A (GlpA) protein family of FAD-dependent anaerobic glycerol-3-phosphate dehydrogenase (residues 31-510). A FAD-dependent oxidoreductase domain DAO (Pfam PF01266, residues 31-406) light green box and, a Bacterioferritin-associated ferredoxin (BFD)-like [2Fe-2S] binding domain (Fer2_BDF; PF04324, residues 455-510) dark green box. Black box, flavin binding domain (according to Cole et al., J Bacteriol. 1988 170(6):2448-56). Grey box corresponds to a small NADH binding domain within a larger FAD binding domain (Pyr_redox_2, PF07992, residues 587-910). Blue box correspond to a

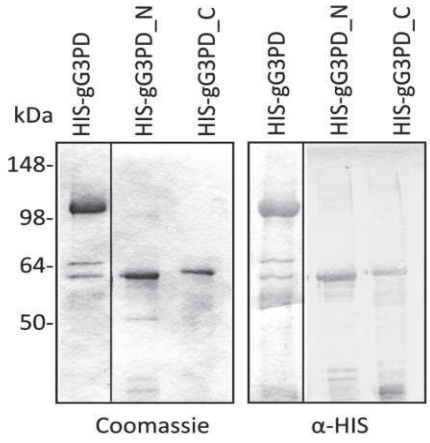
domain found in hypothetical metal-binding proteins (DUF 1667, residues 587-910). Putative 14-3-3 binding site (according to Lalle et al., 2012) are in bold and underlined. Residues common to both the N- and C-terminal half constructs of the gG3PD are in bold and italic. **B) Multiple sequence alignment of the first portion of the G3PD from different species.** Multiple sequence alignment of the first portion (residues 1-510, including the glycerol-3-phosphate dehydrogenase-like domain) of Giardia duodenalis gG3PD (accession number: GL50803_16125) with the identified closer orthologues from Entamoeba histolytica (C4M6H3) and Eggerthella lenta (C8WGG6), and with the Escherichia coli anaerobic-glycerol-3-phosphate dehydrogenase subunit A (GlpA, P0A9C0), the E. coli aerobic glycerol-3-phosphate dehydrogenase (GlpD, P13035), the closely related glycerol-3-phosphate oxidase from Enterococcus faecalis (GlpO, F0PH56) and the human mitochondrial mG3PD (GPD2, P43304). Conserved residues above 30% of identity are highlighted in blue. Darker is the blue, higher is the degree of identity at the position. Orange box, Flavin Adenine Dinucleotide (FAD)-binding motif. Green box, residues that tightly interact (hydrogen bonding and hydrophobic interactions) with FAD and yellow box, residues potentially involved in glycerol-3-phosphate binding (according to Colussi, et al, 2008 and Yeh, et al, 2008).

Supplemental Figure S2



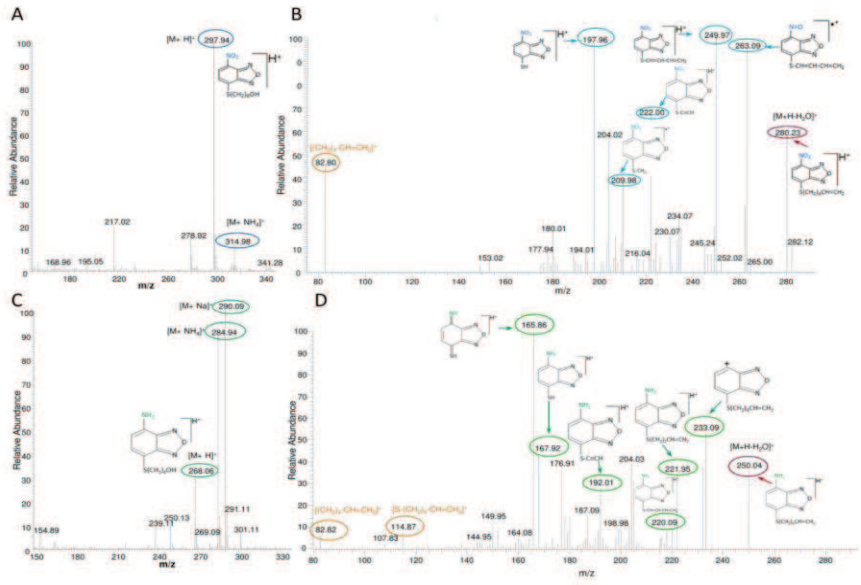
Supplemental Figure S2. Localization of the FLAG-HA-gG3PD during the differentiation stages of *G. duodenalis*. CLSM observations of fixed and permeabilized *G. duodenalis* WB-C6 transgenic parasites expressing the FLAG-HA-gG3PD protein. Trophozoite (panel a, Throph.), encysting parasite after 12h of encystation (panel b and c, Encyst.) and cyst (panel d, Cyst) stained with mouse α-HA mAb (green) and rabbit polyclonal α-g14-3-3 (red). Cyst wall and encystation specific vesicles (ESVs) were stained with Cy3-conjugated α-CWP mAb (grey). Nuclei were stained with DAPI (blue). Displayed micrographs correspond to a single z-stack. T, transmission light acquisition. Scale bars, 5µm.

Supplemental Figure S3



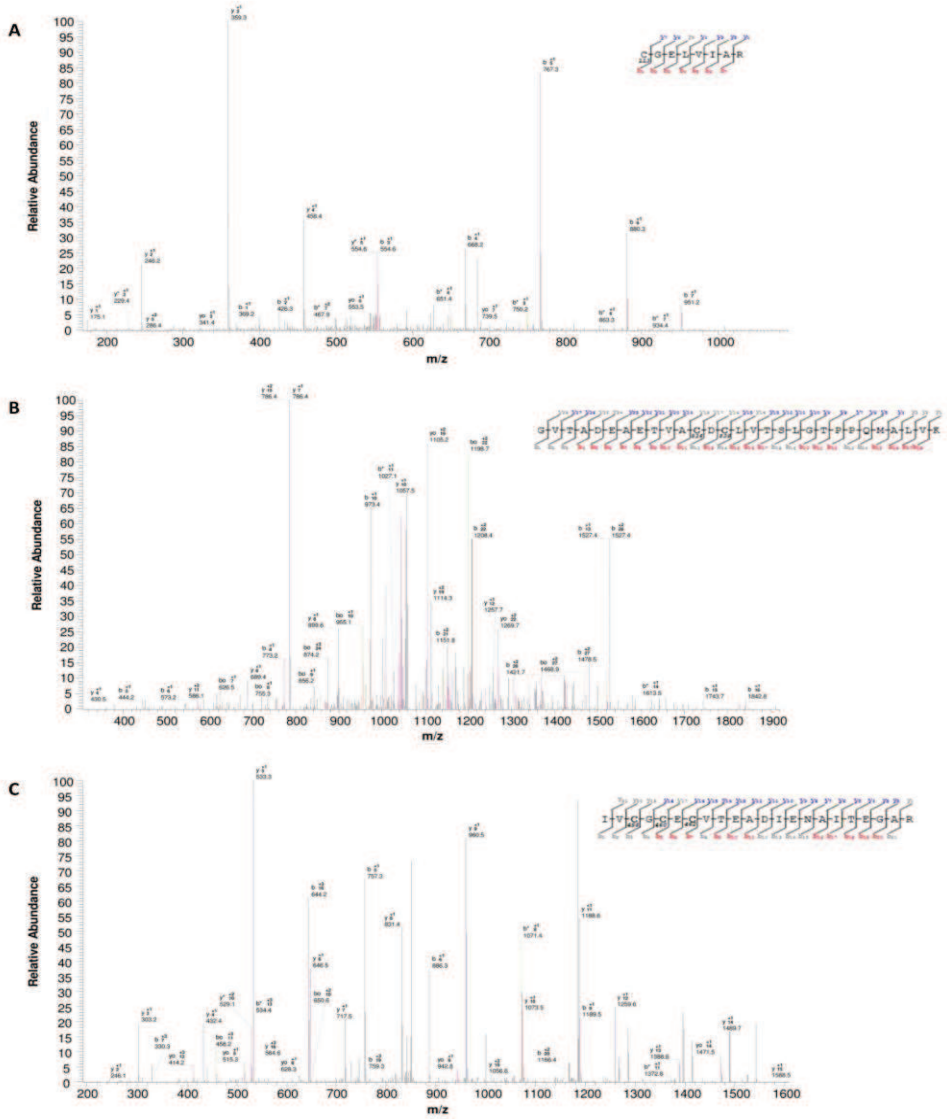
Supplemental Figure S3. Expression of recombinant HIS-gG3PD and deletion mutants in *Escherichia coli*. An aliquot of nickel-affinity purified recombinant HIS-gG3PD (2 mg), HIS-gG3PD_N (0.5 mg) and HIS-gG3PD_C (0.5 mg) were separated on 4-12% SDS-PAGE and either stained with Coomassie blue (Coomassie, left panel) or elettroblotted on nitrocellulose membrane and probed with mouse α-HIS mAb (right panel). Molecular size markers are on the left.

Supplemental Figure S4



Supplemental Figure S4 NBDHEX reduction by sodium dithionite. Full MS of the untreated NBDHEX (A) or NBDHEX treated with sodium dithionite (C). MS/MS of the ions at m/z 298 and 268 (unmodified and reduced NBDHEX, respectively) give rise to neutral loss of a water molecule generating the ions at m/z 280 and 250, whose further fragmentation (MS3) is shown in panel B and D, respectively.

Supplemental Figure S5



Supplemental Figure S5. MS analysis of the in vitro reaction between HIS-gG3PD and NBDHEX. MS/MS spectra matching the gG3PD peptides (113-120) (A), (859-872) (B), and (456-477) (C) carrying cysteine adducts with a mass increase of 265 Da (A and B) and 281 Da (C). The mass shift +265 Da is compatible with an NBDHEX derived adduct in which the nitro group was reduced to an amine one, while the mass increment of 281 Da matches with an NBHEX adduct in which the nitro group was semi-reduced to an hydroxyamine group. In the peptide (113-120) the unique cysteine residue (Cys133) is involved in the adduct as clearly demonstrated by the fragment pattern detected in the MS/MS spectrum (A). The peptide (859-872) contains two cysteine residues (Cys836 and Cys838) and the adduct is more probably located on the first cysteine residue (B). In the peptide (456-477) there are three candidate cysteine residues at positions 458, 460, and 462: the Cys 458 is most likely the one involved in the adduct formation, but the involvement of the second or the third cysteine cannot be excluded (C).

References

1. Colussi, T., Parsonage, D., Boles, W., Matsuoka, T., Mallett, T.C., Karplus, P.A., *et al.* (2008). Structure of alpha-glycerophosphate oxidase from Streptococcus sp.: a template for the mitochondrial alpha-glycerophosphate dehydrogenase. *Biochemistry*. 47, 965-977.

PART III

1 - *Plasmodium berghei* and *Plasmodium knowlesi* phylogenetic profiling.

As described in Part I, we reconstructed and analyzed *P. falciparum* phylogenetic profiles by using the reference set RS2 and the distance correlation as a measure profile similarity. In order to perform comparative analyses of *Plasmodia* interactomes, we applied the same procedure to construct phylogenetic profiles of *P. knowlesi* and *P. berghei*. These two parasites infect primates (*P. knowlesi*) and rodents (*P. berghei*) and are considered *in vivo* models for human malaria.

Although the genome sequences are available for both the parasites, the status of the gene-annotation is still incomplete making difficult the assessment procedure based on the construction of the gold standard data sets. In fact, information related to the two parasites in KEGG database results in a limited number of TPs, especially in the case of *P. knowlesi*.

Considering that the annotation of *P. falciparum* proteins deeply overcomes that one of other *Plasmodia* species, we exploited *P. falciparum* gold standards to derive the corresponding ones of *P. berghei* and *P. knowlesi*. In order to approach this, we used OrthoMCL database (www.orthomcl.org) where groups of orthologous proteins are collected. On these bases, the process of gold standard construction for these *Plasmodium* was done converting the protein pairs in gold standards of *P. falciparum* into the corresponding pairs of orthologous of *P. berghei* and *P. knowlesi*.

The two sets of gold standards (see Table I for details) were used to assess the predicted interactions by phylogenetic profiling of *Plasmodium* parasites.

In figure 1 the predictive performances obtained for *P. berghei* (panel A) and *P. knowlesi* (panel B) are shown as ROC curves.

Table I: summary of the numbers of TP and TN obtained for each set of gold standards

	GS-fun		GS-fun_OGS	
	TP	TN	TP	TN
P. berghei	15320	93395	16180	22207
P. knowlesi	1873	4087	16024	23564

In both cases, the best results were obtained by using gold standards derived from *P. falciparum* orthologs (green lines). In particular, in the case of *P.*

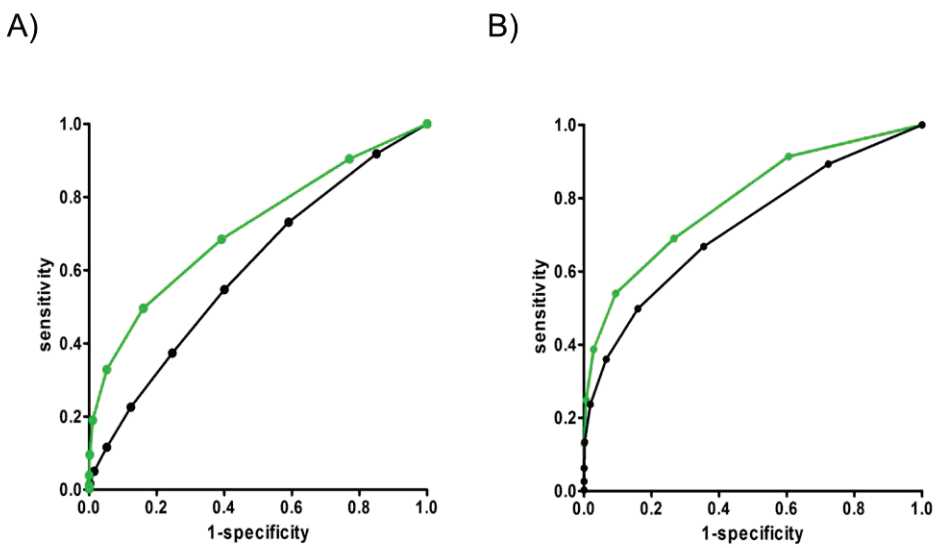


Figure 1. Evaluation of the prediction performances of *P. berghei* and *P. knowlesi* data set using KEGG derived and the gold standards derived from *P. falciparum* ones. *P. berghei* data on panel A, and *P. knowlesi* data on panel B. The best predictive performance is obtained using the gold standards obtained from *P. falciparum* ones (GS-fun_GSO; green lines) in both cases, while the performance assessed by using the GS-fun gets worse results (black lines).

berghei, the AUC obtained by using these gold standards is equal to 0.712, while the AUC obtained using GS-fun$_{Pb}$ is equal to 0.609. For *P. knowlesi*, the AUC obtained using GS-fun_GSO$_{Pk}$ is equal to 0.794, while the AUC obtained using GS-fun$_{Pk}$ is equal to 0.721.

2 – Phylogenetic profiling to predict host-parasite interactions in *P. falciparum* and *P. knowlesi*.

In order to look for novel targets of drug or vaccine development, a big challenge in malaria research is to identify the PPI networks taking place between the host and the parasite. However, in the case of the infected erythrocytes, usually used experimental and computational approaches based on the transcriptomic analyses of the host-parasite system under physiological and perturbed conditions are hampered by the absence of the transcriptional activity of the erythrocytes.

Considering that the similarity of the phylogenetic profiles indicates that two interacting proteins need to co-evolve to perform their biological function, we applied the phylogenetic profile strategy to predict "hybrid" PPI networks of the host-parasite system.

We applied this interspecies phylogenetic profile method to the two host-parasite systems: *i) P. falciparum* and the human erythrocyte (PfH) and *ii) P. knowlesi* and the erythrocyte of its natural host Rhesus monkey (*Macaca mulatta*) (PkM).

In order to approach this, we exploited mass spectrometry data available for the human and monkey erythrocyte proteomes, and then we derived the phylogenetic profiles of the corresponding proteins by using RS2 (Reference Set 2, see paragraph 4.1 in Part I). The distance correlation was adopted to measure the similarity between *Plasmodium*-*Plasmodium*, *Plasmodium*-erythrocyte, and erythrocyte-erythrocyte phylogenetic profiles.

A threshold of the distance correlation equal to 0.6 was established, resulting in the selection of 2211 proteins (1933 and 278 from the parasite and the host

respectively) involved in 53937 interactions for the PfH system and 2286 proteins (1509 and 777 from the parasite and the host respectively) with 204966 interactions for the PkM system. In figure 2 the predicted Plasmodium-erythrocyte interactomes are shown, where full nodes are the *Plasmodium* proteins, while red and green nodes represent human and monkey erythrocyte proteins respectively.

Further analyses will be carried out to explore the two PPI networks and to remove FPs due to the trivial connections between orthologs in the two organisms. This should highlight functionally significant host-parasite interactions that will be the object of experimental validations.

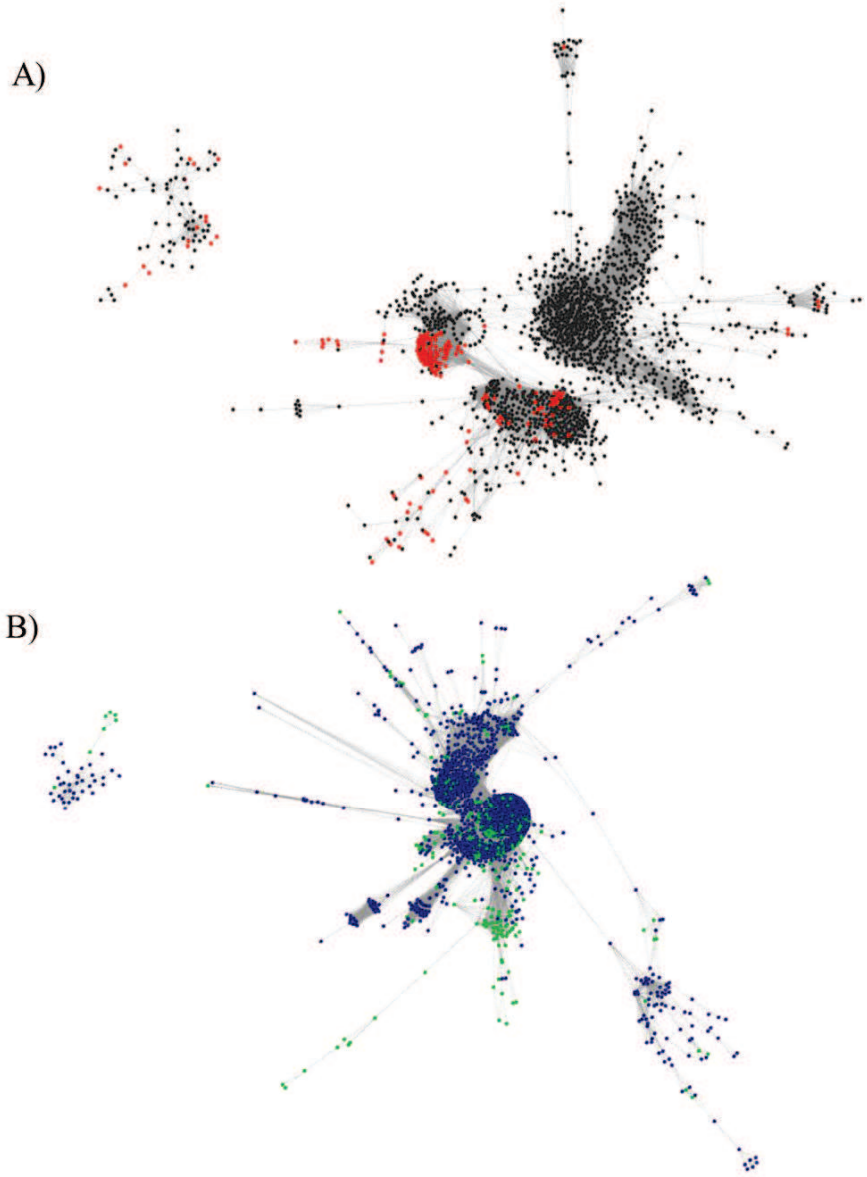


Figure 2. "Hybrid" interactomes of host-parasite systems. A) The main two clusters of the "hybrid" interactome of *P. falciparum* proteins with the human erythrocyte proteins are represented. Black nodes correspond to *P. falciparum* proteins, while red nodes represent the human erythrocyte proteins. The edges are the interactions. B) The main two clusters of the "hybrid" interactome of *P. knowlesi* and Rhesus monkey erythrocyte proteins are represented. Blue nodes correspond to *P. knowlesi* proteins, while green nodes represent the erythrocyte proteins of Rhesus monkey.

3 - A multidomain protein from *Giardia duodenalis*, a rosetta stone case.

Beside the inference of PPIs, the rosetta stone method can be utilized to reconstruct the evolution of gene fusion events exploiting "rosetta-stone proteins" phylogenesis.

In order to explore this approach, we focused on a specific case of study, that is a protein containing a glycerol-3-phosphate dehydrogenase (gG3PD) domain in *Giarda duodenalis*. *G. duodenalis* is a protozoan parasite, which causes giardiasis, the most common feedborne and waterborne human gastroeneteric disease (Ryan & Cacciò, 2013; Halliez & Buret, 2013). Both energy and intermediate metabolism of this parasite have been shown to represent potential targets for drug screening and design (Müller and Hemphill, 2013; Watkins and Eckmann 2014). As gG3PD is a key enzyme at the crossroad of glycolysis, redox and fatty acid metabolism, it represents a potential candidate for drug development.

Preliminary BLAST data base searches suggested that gG3PD is formed by three different domains (a, b, c), and thus can be a potential "rosetta stone protein".

In order to reconstruct its phylogenesis, we developed a specific database search strategy. To avoid any kind of bias possibly introduced by redundancies and/or low quality sequences, we used RS1, the largest, non redundant set among the high quality genome reference sets that we obtained (see paragraph 4.1 in Part I). We then performed five different FASTA searches against the 774 genomes in RS1, using as queries: the entire $gG3PD_{abc}$, and four virtual peptides, the $gG3PD_{ab,}$ formed by the first two domains (a: 1- 500 aa; b: 501-

950 aa), and each of the three domains, $gG3PD_a$, $gG3PD_b$ and $gG3PD_c$ (c: 951-1111 aa).

The Smith-Waterman algorithm, available in the FASTA package (version 36; Pearson & Lipman, 1988), was utilized. A match between sequences was considered positive when the corresponding alignment has a coverage higher than 70%, and an E-value lower than 10^{-6}. In the case of the smallest virtual peptide $gG3PD_c$, the E-value threshold was set at 10^{-3}.

We selected five groups of protein sequences matching the criteria and investigated the evolutionary relationships within each group.

Results clearly show that $gG3PD_a$ and $gG3PD_b$ are closely related to bacterial G3PD domains and FAD/NAD binding domains respectively. We found only 33 matches between $gG3PD_c$ and bacterial proteins, whose annotations suggest a "metal-binding" function. Interestingly, only one match was obtained for $gG3PD_{abc}$ with a protein from the protozoan *Entamoeba hystolitica* that causes intestinal infectious disease in humans and only one match for the $gG3PD_{ab}$ with a protein, from *Eggerthella lenta* that is part of the human colon bacterial flora.

The inspection of the Tree of Life revealed that *E. hystolitica* and *G. duodenalis* not only share the same niche in the human host, but they have also a common origin in that they form a single branch in the Protozoa clade.

Altogether, these results provide highlights on the origin of gG3PD as a "rosetta-stone protein". A possible scenario is the one shown in figure 3, a gene fusion event occurred between $gG3PD_a$ and $gG3PD_b$ in *E. lenta,* then the fused gene was acquired by an eukaryote sharing the same ecological niche in the human intestine and that can be considered as the common ancestor of *E. hystolitica* and *G. duodenalis*.

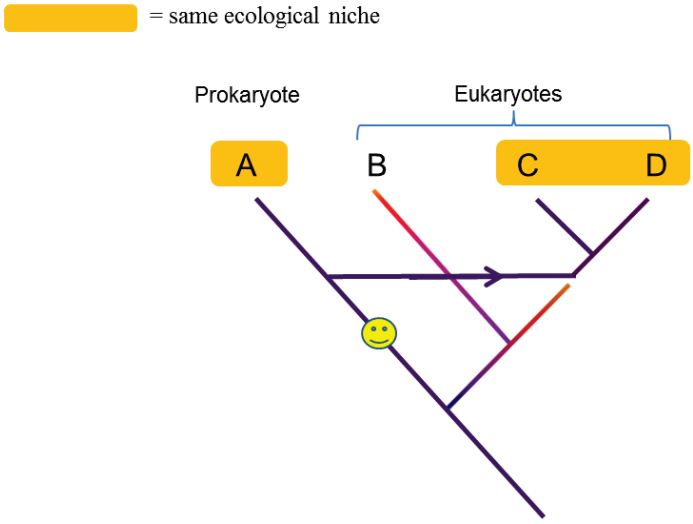


Figure 3. A possible scenario of gG3PD origin. A fusion event (yellow smiling face) took place in a prokaryote (A) followed by a lateral gene transfer to the ancestor of the eukaryotes C and D.