Phylogenomic analyses of malaria parasites and evolution of their exported proteins

Christian Pick1†, Ingo Ebersberger2†, Tobias Spielmann3, Iris Bruchhaus3 and Thorsten Burmester1*

Abstract

Background: Plasmodium falciparum is the most malignant agent of human malaria. It belongs to the taxon Laverania, which includes other ape-infecting Plasmodium species. The origin of the Laverania is still debated. P. falciparum exports pathogenicity-related proteins into the host cell using the Plasmodium export element (PEXEL). Predictions based on the presence of a PEXEL motif suggest that more than 300 proteins are exported by P. falciparum, while there are many fewer exported proteins in non-Laverania.

Results: A whole-genome approach was applied to resolve the phylogeny of eight Plasmodium species and four outgroup taxa. By using 218 orthologous proteins we received unanimous support for a sister group position of Laverania and avian malaria parasites. This observation was corroborated by the analyses of 28 exported proteins with orthologs present in all Plasmodium species. Most interestingly, several deviations from the P. falciparum PEXEL motif were found to be present in the orthologous sequences of non-Laverania.

Conclusion: Our phylogenomic analyses strongly support the hypotheses that the Laverania have been founded by a single Plasmodium species switching from birds to African great apes or vice versa. The deviations from the canonical PEXEL motif in orthologs may explain the comparably low number of exported proteins that have been predicted in non-Laverania.

Background

Malaria is one of the most common infectious diseases, putting about two billion humans at risk and resulting in about one million fatalities each year [1]. Malaria is caused by protozoan parasites of the genus Plasmodium (Haemosporidae; Apicomplexa). Species of this genus undergo a complex life cycle including an asexual proliferation phase in the erythrocytes of vertebrate hosts.

Although hundreds of Plasmodium species are currently known, only few infect humans. In moderate climate zones, human malaria infection is largely due to P. vivax, but the life-threatening form of this disease is almost exclusively caused by P. falciparum. About 60 years ago, the high pathogenicity of P. falciparum led to the proposal that this parasite may be a rather recent acquisition from a non-human host [2]. Since then, it has become evident that P. falciparum indeed is closely related to other Plasmodium species from African great apes [3,4]. Together they constitute the subgenus Laverania and several reciprocal host switches have occurred during the evolution of this group of malaria parasites [5-9].

The evolutionary ancestry of P. falciparum and the other Laverania is still a matter of debate. Until now, it has not been conclusively agreed on whether this subgenus is more closely related to other mammalian malaria parasites or whether it shares a common ancestry with bird-infecting Plasmodium species (reviewed in [10]). Most molecular phylogenetic studies of the genus Plasmodium are based on the analysis of single proteins such as cytochrome b oxidase, adenylosuccinate lyase, and caseinolytic protease C [10]. While these proteins contain sufficient phylogenetic information to resolve the relationships within the Laverania, multiple substitutions per site (homoplasy) limit their utility at a deeper phylogenetic level [10].

Upon invasion by P. falciparum, erythrocytes are subjected to an extensive remodeling process resulting in altered mechanical and adhesive properties [11].
Prominent examples include the formation of cytoadherence knobs at the erythrocyte membrane and the associated exposure of PfEMP1 (P. falciparum erythrocyte membrane protein) at the surface of the infected cell [12]. Plasmodium proteins involved in this remodeling process have to pass the parasitophorous vacuole membrane (PVM) on their way from the parasite into the erythrocyte; most of these proteins are characterized by a hydrophobic signal sequence for targeting the protein to the endoplasmic reticulum (ER) and a sequence motif (RxLxE/Q/D) either referred to as Plasmodium export element (PEXEL; [13]) or vacuolar transport signal [14]. The PEXEL motif is cleaved by the aspartyl protease plasmspein V in the ER of the parasite [15-17] and the nascent protein is released into the parasitophorous vacuole. From there it is transported through the PVM into the host cell by the Plasmodium translocon of exported proteins (PTEX; [18]). Predictions based on the presence of the PEXEL suggest that more than 300 P. falciparum proteins are exported into the host cell [13,14,19,20]. Notably, PfEMP1s are structurally different, having an export element that precedes the signal sequence (RxLxE/Q/D) either referred to as PEXEL; [13]) or vacuolar transport signal [14].

The conservation of plasmspein V and the components of PTEX throughout the genus Plasmodium indicates that the same protein export machinery is used by all Plasmodium species [16-18]. In addition, PEXEL sequences from P. falciparum proteins proved to be functional in rodent malaria parasites [13] and vice versa [22]. Thus, in principle, the screens to detect exported proteins in P. falciparum should be extendable to other Plasmodium species. However, surprisingly few proteins have been detected outside of the Laverania using the P. falciparum PEXEL motif, and it has been suggested that these species export fewer proteins into the host cell than P. falciparum [13,14,19]. Non-Laverania, however, also induce elaborate morphological changes in their host cells and the low number of predicted exported proteins may argue for a prominent role of PEXEL-negative exported proteins (PNEPs; reviewed in [23]). An additional, thus far unexplored, explanation could be a slightly different consensys of the PEXEL motif in Plasmodium taxa other than Laverania that could hamper the prediction of these proteins. This would inevitably lead to an underestimation of the respective exportomes.

Here, we took advantage of the available genomic sequences from eight Plasmodium species and four other apicomplexan species. Orthologous proteins were identified and used (i) to reconstruct the phylogeny of these species, (ii) to obtain a set of exported P. falciparum proteins that are conserved throughout Plasmodium evolution, and (iii) to investigate the evolutionary plasticity of the corresponding Plasmodium export elements.

Methods
Source of sequence data
The genomic sequences of P. falciparum 3D7 [24], P. yoelii 17XNL [25], P. berghei Anka [26], P. chabaudi AS [26], P. knowlesi H [27], P. vivax Sal-1 [28], as well as P. reichenowi and P. gallinaceum (both unpublished data produced by the Wellcome Trust Sanger Institute; used with permission) were obtained from PlasmoDB v. 6.1 [29]; sequences of Toxoplasma gondii ME49 were obtained from ToxoDB v. 5.2 [30], sequences of Babesia bovis T2Bo from Integr8 [31], sequences of Cryptosporidium parvum Iowa from CryptoDB [32]; sequences of Theileria annulata Ankara [33] were downloaded from the Sanger Institute http://www.sanger.ac.uk/resources/downloads/protozoa/.

Collection of orthologs
The dataset of orthologous proteins for phylogeny reconstructions was compiled as described before [34]. In brief, InParanoid-TC was used with P. falciparum, P. vivax, P. knowlesi, P. yoelii, P. berghei, P. chabaudi, T. gondii, and B. bovis as primer taxa. For 921 proteins orthologs were present in all eight primer taxa. These 921 core orthologs served then as input for HaMStR to search for the corresponding proteins in P. reichenowi, P. gallinaceum, C. parvum, and T. annulata. Following search species - reference species pairs were used in the HaMStR search: P. reichenowi - P. falciparum, P. gallinaceum - P. falciparum, C. parvum - T. gondii, and T. annulata - B. bovis. HaMStR could extend 218 core orthologs with sequences from all four species such that each ortholog group consisted of twelve sequences. The amino acid sequences for each of the 218 core orthologs were aligned with MAFFT [35] using the options -maxiterate 1000 and -localpair. The 218 single alignments were concatenated to form a super-alignment with 192,102 aa positions. This super-alignment was processed twice: (i) positions for which less than half of the sequences were represented by an amino acid were removed, and (ii) Gblocks 0.91b [36] was applied using the following parameters: -minimum number of sequences for a conserved position was set to 7; -minimum number of sequences for a flanking position was set to 10; -maximum number of contiguous nonconserved positions was set to 4; -minimum length of a block was set to 10; and -allowed gap positions was set to none. To obtain the collection of exported proteins that have functionally equivalent orthologs in the other Plasmodium species, the two most comprehensive
predictions of exported \textit{P. falciparum} proteins were used \cite{19,20}. These predictions contain 396 and 422 proteins (not including the structurally distinct \textit{PIEMP1}s), respectively; the combination of both resulted in a non-redundant set of 531 putatively exported proteins. Each protein was used as query for a tBLASTn search against the \textit{P. falciparum} genome. Proteins, for which the \( E \) value of the best BLAST hit (not considering the hit against itself) was larger than \( 10^{-10} \), were considered to have no paralogs present in \textit{P. falciparum} and were used for further analysis. For each paralog-free protein a reciprocal tBLASTn search was performed to identify candidate orthologs in the other \textit{Plasmodium} species (\( E \) value cut-off: \(< 10^{-10} \)). Proteins with a single ortholog present in each of the eight \textit{Plasmodium} species were aligned with MAFFT \cite{35} using the options \texttt{--maxiterate 1000} and \texttt{--localpair}.

\textbf{Phylogeny reconstruction}

Maximum likelihood (ML) trees were reconstructed with RAxML v. 7.2.2 \cite{37} using the WAG model \cite{38} of amino acid sequence evolution with empirical amino acid frequencies (option \texttt{F}). Substitution rate heterogeneity was modeled using a gamma distribution, allowing for a fraction of invariant sites (option \texttt{GAMMA}). Bayesian tree search was performed with PhyloBayes v. 3.2 \cite{39} using the WAG model. Four MCMC chains were run for 10,000 cycles. Every 10\(^{th}\) cycle was sampled and convergence of the chains was pair-wise checked with \texttt{bpcomp} allowing for a burn-in of 1,000 cycles. Increasing the burn-in or usage of other models of amino acid sequence evolution such as the CAT \cite{40} or LG model \cite{41} did not change the results (not shown).

\textbf{Testing of alternative phylogenies}

The small number of taxa in our study allows the evaluation of every possible tree topology. However, we reduced the number of tested trees by imposing the following constraints: monophyly of the genus \textit{Plasmodium}; monophyly of \textit{B. bovis} and \textit{T. annulata} (Piroplasmsida); monophyly of \textit{T. gondii} and \textit{C. parvum} (Eimeriorina); monophyly of \textit{P. yoelii} and \textit{P. berghei}; monophyly of \textit{P. falciparum} and \textit{P. reichenowi}; monophyly of \textit{P. vivax} and \textit{P. knowlesi}. Note that all seven constraints represent accepted evolutionary relationships (see references in Table 1) except the monophyly of \textit{T. gondii} and \textit{C. parvum} \cite{42}, and have been confirmed by our unrestricted heuristic tree searches. We computed the likelihood of the resulting 105 alternative tree topologies with TREE-PUZZLE v. 5.2.pl21.1 \cite{43} using the WAG model of sequence evolution. Substitution rate heterogeneity was modeled with a gamma distribution assuming four rate categories and empirical amino acid frequencies. Hypothesis testing was performed using the routines provided by TREE-PUZZLE and by CONSEL \cite{44}.

\textbf{Sequence analysis}

Pairwise amino acid identities and similarities were calculated with GeneDoc v. 2.6 \cite{45} using the Blosum 62 model. PEXEL sequences of the \textit{P. falciparum} proteins were identified via a match to the published consensus sequences \cite{13,14,19,20}. The putative PEXEL sequences of proteins from other \textit{Plasmodium} species were extracted by aligning these proteins to their ortholog in \textit{P. falciparum}; we then used the homologous amino acid positions to the \textit{P. falciparum} PEXEL as candidate export elements in these species. PEXEL sequences from the individual proteins were aligned separately for each species by hand and the corresponding PEXEL motifs were generated with WebLogo \cite{46}. Presence of hydrophobic signal sequences was assessed using SignalP v. 3.0 \cite{47}.

\textbf{Results and Discussion}

\textbf{Evolutionary ancestry of \textit{P. falciparum} and other \textit{Laverania}}

We extracted the genomic sequences of eight \textit{Plasmodium} species (\textit{P. falciparum}, \textit{P. reichenowi}, \textit{P. vivax}, \textit{P. knowlesi}, \textit{P. gallinaceum}, \textit{P. chabaudi}, \textit{P. yoelii}, and \textit{P. berghei}) and four additional apicomplexan species (\textit{T. gondii}, \textit{C. parvum}, \textit{T. annulata}, and \textit{B. bovis}) from public databases. HaMStR, a Hidden Markov Model based tool \cite{34}, was used to identify 218 proteins with orthologs in all twelve species (Additional file 1). This number is similar to that used in a recent phylogenomic study of eight apicomplexan species, including two species from the genus \textit{Plasmodium} \cite{48}.

The single alignments of the 218 proteins were concatenated and positions for which less than half of the taxa were represented by an amino acid were removed. This resulted in a super-alignment with 135,360 aa positions (Additional file 2), which was used for initial maximum likelihood (ML) and Bayesian tree reconstructions. While tree topologies inferred from the ML analysis were identical with those obtained in later analyses (see below; Figure 1), MCMC chains did not converge on a single topology, indicating that the dataset includes conflicting phylogenetic information. Therefore, the 135,360 aa alignment was further processed using Gblocks \cite{36}. This procedure has been demonstrated to improve phylogenetic analyses by reducing the impact of misaligned regions (due to very high sequence divergence) and homoplasy (due to sequence saturation) \cite{49}. We obtained a final alignment comprising 49,521 aa positions and no missing data (Additional file 3). In Bayesian analysis, MCMC chains readily converged on the same topology (maxdiff: 0; meandiff: 0). In the resulting
consensus tree all nodes received strong support and the same topology was obtained in ML analyses (Figure 1).

The phylogenetic analyses show that the eight Plasmodium species form a monophyletic clade (100% bootstrap support and 1.00 Bayesian posterior probabilities). The malaria parasites from rodents (P. chabaudi, P. yoelii, and P. berghei) are clearly separated from those infecting birds and primates (100% bootstrap support and 1.00 Bayesian posterior probabilities). Notably, the Laverania (P. falciparum and P. reichenowi) do not group with the other primate-infesting malaria parasites, but form a well-supported clade with P. gallinaceum (99% bootstrap support and 1.00 Bayesian posterior probabilities).

ML as well as Bayesian methods return only the best tree and thus provide no information on other tree topologies with likelihoods that may not be significantly worse. To address this issue, 104 alternative tree topologies were tested by inferring their expected likelihood weights (ELW; [50]) and their probabilities in the approximately unbiased (AU) test [51]. All alternative tree topologies (including those with P. gallinaceum being the sister group of mammal Plasmodium species) were rejected with high confidence (Figure 2; Additional Table 1 Molecular phylogenetic analyses attempting to resolve the relationships among malaria parasites

| Reference                | Gene*          | Outgroup taxa                        | Number of ingroup taxa | Laverania + avian malaria parasites |
|--------------------------|----------------|--------------------------------------|------------------------|-------------------------------------|
| Waters et al. 1991 [53]  | SSU rRNA       | Acanthamoeba castellani              | 10                     | Yes                                 |
| Escalante and Ayala 1994 | SSU rRNA       | Theileria annulata Babesia bovis     | 11                     | No                                  |
| Escalante and Ayala 1995 | SSU rRNA       | Colpidium campylum Euplotes aediculatus Glaucoma chattoni Opisthohecta hennegyui Paramecium tetraurelia | 5                      | Yes                                 |
| Qari et al. 1996 [68]    | SSU rRNA       | Toxoplasma gondii, Paramecium tetraurelia | 13                     | No                                  |
| Escalante et al. 1998    | Cyt b          | Toxoplasma gondii                    | 8                      | No                                  |
| Rathore et al. 2001      | SSU rRNA       | Toxoplasma gondii                    | 8                      | Yes                                 |
| Perkins and Schall 2002  | Cyt b          | Theileria annulata Leucocytaezoon dubreuli Leucocytaezoon simondi | 52                     | No                                  |
| Kissinger et al. 2002    | SSU rRNA       | Theileria annulata Babesia equis     | 10                     | Yes                                 |
| Leclerc et al. 2004      | SSU rRNA       | Toxoplasma gondii Sarcozystis fusiformis Babesia bovis | 21                     | No                                  |
| Roy and Irimia 2008      | SSU rRNA       | Leucocytaezoon caulleryi Leucocytaezoon sabrazesi | 18                     | No                                  |
| Martinsen et al. 2008    | Cyt b, Cox I, ClpC, ASL (concatenated) | Leucocytaezoon spp. | 57                     | No                                  |
| Ollomo et al. 2009 [4]   | Cyt b, Cox I, Cox III (concatenated) | Leucocytaezoon caulleryi Leucocytaezoon sabrazesi | 17                     | No                                  |
| Krief et al. 2010 [7]    | Dhfr-ts, Msp2 (concatenated) | Leucocytaezoon caulleryi Leucocytaezoon sabrazesi | 42                     | No                                  |
| Silva et al. 2010 [53]   | 29 proteins (concatenated) | Theileria annulata Theileria annulata | 8                      | No                                  |

Note that only analyses are listed that included non-Plasmodium species as an outgroup taxa

*Abbreviations: SSU rRNA, 18S small subunit ribosomal RNA; Cyt b, cytochrome b; ClpC, caseinolytic protease C; Cox I and III, cytochrome oxidase I and III; ASL, adenylosuccinate lyase; Dhfr-ts, dihydrofolate reductase-thymidylate synthase; Msp2, merozoite surface protein 2.

1. Pick et al. BMC Evolutionary Biology 2011, 11:167 http://www.biomedcentral.com/1471-2148/11/167 Page 4 of 12
Thus, the position of *P. gallinaceum* as sister group to the Laverania receives unambiguous support from the data.

Until now, two other whole-genome approaches attempted to resolve the evolutionary relationships of the eight *Plasmodium* species. Dávalos and Perkins [52] based their analyses on a set of 104 proteins (~26,000 aa positions), recovering the same topology among *Plasmodium* species as displayed in Figure 1. However, no outgroup taxa were included to root the tree, and thus no information on the evolutionary ancestry of the Laverania could be provided. Silva et al. [53], on the other hand, based their analyses on a set of 29 proteins (~12,000 aa positions) and used two species from the genus *Theileria* to root the tree. While they proposed the monophyly of mammalian *Plasmodium* species, some of their results supported a grouping of *P. gallinaceum* and the Laverania.

Dávalos and Perkins [52] as well as Silva et al. [53] both used slow-evolving proteins for their phylogenetic inferences. To assess the effect of the evolutionary rate on our analysis, we partitioned the 218 proteins of our dataset. We first computed a ML tree for each protein individually. The length of this protein tree, i.e., the sum of its branch lengths, then served as an approximation for the evolutionary rate (Figure 3; see also Additional file 5). Subsequently, the proteins were categorized into three subsets according to their tree lengths. Dataset 1

![Figure 1 Phylogenetic relationships of eight malaria parasites](image1)

**Figure 1** Phylogenetic relationships of eight malaria parasites
Phylogenetic tree reconstructions were based on 218 proteins. The single alignments were concatenated to form a super-alignment and problematic alignment regions were subsequently removed with Gblocks. This procedure resulted in an alignment with 49,521 aa positions (no missing data), which was used for ML tree reconstruction and Bayesian tree search. *T. gondii, C. parvum, T. annulata*, and *B. bovis* were used as an outgroup to root the tree. Only the ML tree is displayed, but the topology of the Bayesian tree was identical. Numbers at the nodes denote bootstrap support values (left) and Bayesian posterior probabilities (right). The scale bar equals 0.05 expected substitutions per site. See also Additional files 1, 2 and 3.

![Figure 2 Alternative relationships among major Plasmodium lineages resulting from different root placements](image2)

**Figure 2** Alternative relationships among major *Plasmodium* lineages resulting from different root placements. The results for the individual likelihood ratio tests are given below each tree. ELW, expected likelihood weights [50]; AU, approximately unbiased test [51]. *T. gondii, C. parvum, T. annulata*, and *B. bovis* were used as an outgroup. Note that only the results from analyses with rate heterogeneity are shown; results from analyses without rate heterogeneity were essentially the same. See also Additional file 4.

![Figure 3 Evolutionary rates of the 218 proteins used for phylogenetic inferences](image3)

**Figure 3** Evolutionary rates of the 218 proteins used for phylogenetic inferences. The length of the individual protein trees, i.e., the sum of its branch lengths, served as an approximation for the evolutionary rate. Note that only the proteins of primer taxa (*P. falciparum, P. vivax, P. knowlesi, P. yoelii, P. berghei, P. chabaudi, T. gondii, and B. bovis*) were considered. See also Additional file 5.
comprised 65 slow-evolving proteins (tree lengths of less than two expected substitutions per site). Dataset 2 comprised 88 proteins evolving at an intermediate speed (tree lengths of two or more but less than four expected substitutions per site). Dataset 3 comprised the 66 fast-evolving proteins (tree length of four or more expected substitutions per site). For subsequent tree reconstruction, 65 proteins were randomly chosen from each partition such that the same number of proteins was used for each dataset. The individual alignments were concatenated, processed with Gblocks as described and used for ML tree reconstruction (Figure 4). All three datasets agree in placing *P. gallinaceum* as sister of the *Laverania*. The topology of the tree was thus identical with that inferred from the complete dataset (Figure 1). We conclude that our reconstruction of the *Plasmodium* phylogeny does not depend on the evolutionary rates of the proteins used for the phylogeny.

The bootstrap support for the clade consisting of *P. gallinaceum* and *Laverania* was maximal for the dataset comprising proteins evolving at an intermediate speed (98%) and minimal for the dataset comprising the fast-evolving proteins (76%). The branch leading to the clade consisting of *P. gallinaceum* and *Laverania* was short (~0.02 expected substitutions per site; cf. Figure 4). When using fast-evolving proteins, multiple substitutions in the dataset might confound the phylogenetic signal leading to artifacts due to long branch attraction [54]. On the other hand, using only slow-evolving proteins is likely to result in a dataset with a phylogenetic signal that is too weak to resolve this branch (see also Additional file 6). This may explain why proteins evolving with an intermediate rate provide the most robust tree.

The finding of a relationship between the *Laverania* and avian malaria parasites agrees with earlier studies by Waters et al. [55], Escalante and Ayala [56], and Kissinger et al. [57]. However, it contradicts more recent results by Perkins and Schall [58], Leclerc et al. [59], Roy and Irimia [60] and Martinsen et al. [61]. This discrepancy may be attributed to the limited phylogenetic information in the few proteins that were used in those studies [10]. While the selection of proteins may have some effect (see above), the number and choice of the outgroup taxa deserve particular attention (e.g., [62]). Alternative root placements lead to different conclusions about the order in which the individual *Plasmodium* species emerged (c.f. Figure 2). In many previous studies, only a single outgroup taxon was used (Table 1). Moreover, in some cases this outgroup was evolutionarily so distantly related that a meaningful placement of the root is unlikely (e.g., [63]). Most recent studies of *Plasmodium* phylogeny used selected species from the closely related genus *Leucocytozoan* as an outgroup (c.f. Table 1). However, the limited amount of sequence data available for this taxon - mainly a few mitochondrial genes - currently prevents its use in phylogenomic studies. Other haemosporidians (i.e., species from the genera *Haemoproteus*, *Parahaemoproteus*, and *Hepatocystis*) should not be considered as an outgroup since the genus *Plasmodium* has been shown to be paraphyletic with respect to these taxa (e.g., [61]). Alternative strategies for a reliable root placement employ the inclusion of multiple outgroup taxa to break the branch separating the ingroup from the outgroup, and the use of a comprehensive set of proteins [64]. Our trees include four apicomplexan species as an outgroup and are based on 218 orthologous proteins. We have obtained identical
tree topologies by employing different tree reconstruction methods (ML and Bayesian inference) and different models of sequence evolution. Moreover, our findings remain unchanged when we use proteins with different evolutionary rates. Ultimately, likelihood ratio tests rejected all alternative tree topologies. Thus, we are confident that our root placement is robust and that \textit{P. gallinaceum} and the Laverania indeed share a common ancestry.

An avian parasite as sister to the Laverania has significant implications: it suggests that a host switch from birds to African great apes or vice versa has occurred. Host switches have repeatedly taken place during the evolution of avian \textit{Plasmodium} species [61]. Moreover, avian \textit{Plasmodium} species are able to infect mammals under experimental conditions [65]. Both observations are congruent with an evolutionary scenario in which the laveranian lineage was established by a single \textit{Plasmodium} species switching from birds to African great apes. Subsequent diversification of Laverania associated with multiple host switches within the apes eventually led to the emergence of \textit{P. falciparum} in humans [5-9]. Note that this scenario also implies that the great diversity of malaria parasites infecting birds [61] may in fact derive from an early host switch by another mammalian \textit{Plasmodium} species. At present, however, we cannot exclude the alternative scenario in which the avian \textit{Plasmodium} lineage was established by a \textit{Plasmodium} species from the laveranian lineage. Therefore, phylogenomic analyses considering additional \textit{Plasmodium} species (and in particular those infecting birds and squamate reptiles) will be necessary to provide a more detailed picture of how the Laverania emerged.

**Evolutionary plasticity of the \textit{Plasmodium} export element**

The availability of \textit{Plasmodium} genome sequences together with the reliable reconstruction of their phylogenetic relationships provides a robust framework to investigate the evolutionary history of exported \textit{P. falciparum} proteins. Here, we used 531 \textit{P. falciparum} proteins that had been predicted to be exported into the host cell [19,20] to identify functionally equivalent orthologs in the other \textit{Plasmodium} species. BLAST searches in the \textit{P. falciparum} genome identified 102 proteins without any recognizable paralog (Additional file 7), whereas the other 429 proteins mainly belong to large gene families such as RIFINs (repetitive interspersed family) and STEVORs (subtelomeric variable open reading frames). These gene families have a complex evolutionary history and have undergone independent lineage-specific diversifications [19]. This indicates that even if homologs of these proteins exist in the other \textit{Plasmodium} species, they do not necessarily share the same function. These proteins were therefore excluded from further analyses.

Subsequent BLAST searches for homologs of the 102 paralog-free proteins in the other \textit{Plasmodium} species identified 33 proteins with a homolog present in each of the species (Table 2). Orthology between the members in the 33 groups of proteins was confirmed by inferring the corresponding sequence trees with a Bayesian approach as described in the Methods section (Figure 5; Additional file 7). Whereas in 27 cases this tree was congruent to the species tree, six sequence trees differed from the species tree in the position of the \textit{P. gallinaceum} sequences. However, subsequent likelihood ratio tests revealed that superimposing the species tree did not lead to significantly worse likelihoods (Additional file 8). The pairwise similarities between the \textit{P. falciparum} proteins and their orthologs in the other \textit{Plasmodium} species are given in Table 2. The orthologs from \textit{P. reichenowi} display the highest degree of similarity. This finding is expected given the sister group relationship of \textit{P. falciparum} and \textit{P. reichenowi}. Among the remaining six non-laveranian taxa the orthologs from \textit{P. gallinaceum} are overall most similar to the \textit{P. falciparum} proteins. This lends further support to our conclusion that the Laverania and \textit{P. gallinaceum} share a common ancestry.

Both reciprocal best BLAST hit searches and phylogenetic tree reconstructions indicate that the proteins in the 33 groups are encoded by genes that remained single copy throughout evolution (one-to-one orthologs). Ample evidence exists that such one-to-one orthologs are functionally equivalent [66]. Therefore, we conclude that if the \textit{P. falciparum} protein is exported, its orthologs in other \textit{Plasmodium} species are exported as well and hence, that these proteins are suitable to assess the evolutionary plasticity of the PEXEL motif. Note that five of these 33 proteins have already been confirmed to be exported in \textit{P. falciparum} using GFP-constructs ([20]; Table 2). However, five proteins appear not to be exported ([19,20]; Table 2); thus they were omitted from further analyses.

The amino acid alignments of the remaining 28 ortholog groups were used to identify the regions that correspond to the \textit{P. falciparum} PEXEL in the sequences from the other species. Subsequently, the candidate PEXEL sequences from all proteins were extracted and aligned separately for each species (Additional file 8). From these alignments the individual PEXEL motifs were determined and compared to those of \textit{P. falciparum} (Figure 6). The motifs were found to be largely similar across the different \textit{Plasmodium} species. However, several deviations from the functionally important amino acids were observed; the amino acids at position 1 and 3 are crucial for an efficient cleavage by plasmepsin V, while the amino acid at position 5 affects the export rate of the nascent protein [67].
The most prominent difference between the Plasmodium species was found for the positively charged amino acid at position 1 of the PEXEL motif. All 28 P. falciparum proteins harbor an arginine (R), whereas about 20% of the proteins from non-Laverania have a lysine (K) at this position. Three lines of evidence indicate that this alternate PEXEL is nevertheless functional: (i) lysine at position 1 of the PEXEL was found in orthologs of those P. falciparum proteins whose export into the host cell has been confirmed (Figure 5), and thus our observation is not restricted to proteins that might have been erroneously predicted as being exported; (ii) recent experimental evidence suggests that the typical cleavage at the leucine (L) at position 3 can occur in proteins containing lysine at position 1 (PFI1780w and MAL3P8.15; [70]); and (iii) a small number of proteins with a lysine at position 1 of the PEXEL have already been predicted to be exported using a Hidden Markov Model based prediction method (21 in P. falciparum, three or less in each of the other Plasmodium species; cf. [19]). Other deviations at position 1 that are less prominent include the presence of histidine (H) in the P. knowlesi and P. vivax orthologs of PFC0435w and of glutamine (Q) in the P. gallinaceum protein that is
orthologous to PFA0210c (Figure 5). Both PFC0435w and PFA0210c belong to the confirmed set of exported proteins in *P. falciparum* [20] and therefore these PEXEL sequences are likely to be functional as well. Position 3, which almost invariably harbors a hydrophobic leucine (L), was also found almost invariant in the orthologs of the confirmed exported *P. falciparum* proteins. However, several orthologs of *P. falciparum* proteins that have not yet been confirmed to be exported have an isoleucine (I) at this position (Figure 6). Position 5, which is considered to be the least conserved position [13,14], was found to be even more variable in the group of confirmed exported *P. falciparum* proteins.

Even though it remains to be demonstrated that these orthologous proteins are cleaved and exported with the same efficiency, these observations suggest that the PEXEL motif is more variable than previously acknowledged. This provides a possible explanation for the small number of exported proteins predicted for some *Plasmodium* species. Taking this plasticity into account will be essential to arrive at a more comprehensive set of exported proteins for all *Plasmodium* species.

**Conclusion**

Our phylogenetic analyses of orthologs deduced from the *Plasmodium* genomes strongly suggests that the
subgenus Laverania was established by a single host switch from birds to African great apes (or vice versa). However, sequences from additional bird-infecting Plasmodium species and the closely related Haemosporida are required to better understand the early evolution of the Laverania. Exported proteins, as identified by the PEXEL motif, play a major role in Plasmodium virulence and facilitate the parasite’s survival in the host cell. Our results suggest that the number of exported proteins is higher in the non-laveranian Plasmodium species than previously assumed. Comprehensive knowledge of their diversity and evolution will help to unravel the emergence of the high pathogenicity of P. falciparum, and may allow the identification of novel targets for malaria therapy.

Acknowledgements
The P. gallinaceum and P. reichenowi sequence data were produced by the parasite genomes group at the Wellcome Trust Sanger Institute and are available from http://www.sanger.ac.uk/resources/downloads/parasites/. We thank Janus Borner for help with the software applications and Tina Koester for helpful discussions. The studies were supported in part by the DFG priority program SPP 1174 Deep Metazoan Phylogeny [HA 1628/9]. We thank two anonymous reviewers and the editor for their helpful comments, and Kathleen Rankin for correction of the language.

Author details
1Institute of Zoology and Zoological Museum, University of Hamburg, Martin-Luther-King-Platz 3, D-20146 Hamburg, Germany. 2Center of Integrative Bioinformatics Vienna, University of Vienna, Medical University of Vienna, University of Veterinary Medicine Vienna, Dr. Bohrgasse 9, 1030 Vienna, Austria. 3Bernhard-Nocht-Institute for Tropical Medicine, Bernhard-Nocht-Str. 74, D-20359 Hamburg, Germany.

Authors’ contributions
All authors conceived the study. CP and IE analyzed the data. All authors drafted the manuscript. All authors read and approved the final version of the manuscript.

Received: 14 February 2011 Accepted: 15 June 2011
Published: 15 June 2011

Additional material

Additional file 1: 218 P. falciparum proteins with orthologs present in P. reichenowi, P. vivax, P. knowlesi, P. gallinaceum, P. chabaudi, P. yoelii, and P. berghei, as well as the four outgroup taxa (T. gondii, C. parvum, T. annulata, and B. bovis).

Additional file 2: Alignment with 135,360 aa positions. Total amount of missing data was 10.2%; amount of missing data per species was 7.1% for T. gondii, 14.8% for C. parvum, 12.2% for T. annulata, 12.5% for B. bovis, 0.5% for P. falciparum, 35.4% for P. reichenowi, 25.4% for P. gallinaceum, 0.7% for P. knowlesi, 0.6% for P. vivax, 0.8% for P. chabaudi, 4.0% for P. yoelii, and 14.2% for P. berghei.

Additional file 3: Alignment with 49,521 aa positions. Total amount of missing data was none.

Additional file 4: Likehood ratio tests for 104 different tree topologies (with and without rate heterogeneity), including their expected likelihood weights (ELW) and probabilities in the approximately unbiased (AU) test. User specified test trees, as well as TREE-PUZZLE and CONSEL output files are given.

Additional file 5: Sum of the expected substitutions per site for each of the 218 proteins used for phylogenetic inference. Note that only the proteins of primer taxa (P. falciparum, P. vivax, P. knowlesi, P. yoelii, P. berghei, P. chabaudi, T. gondii, and B. bovis) were considered.

Additional file 6: Split network analyses of Plasmodium phylogeny, as reconstructed from the 65 slow-evolving proteins in our dataset. The network was calculated with SplitsTree v. 4.11.3 [71], considering only splits with frequencies of at least 10%. Notably, a sister group position of P. gallinaceum to all mammalian Plasmodium species is not supported.

Additional file 7: 102 paralog-free P. falciparum proteins that are exported into the host cell and information on the presence or absence of orthologs in other Plasmodium species.

Additional file 8: Phylogeny of exported P. falciparum proteins with orthologs present in all Plasmodium species; corresponding amino acid alignments and information on the amount of missing data; P. falciparum PEXEL motifs and orthologous sequences of the other Plasmodium species.

Acknowledgements

The P. gallinaceum and P. reichenowi sequence data were produced by the parasite genomics group at the Wellcome Trust Sanger Institute and are available from http://www.sanger.ac.uk/resources/downloads/parasites/. We thank Janus Borner for help with the software applications and Tina Koester for helpful discussions. The studies were supported in part by the DFG priority program SPP 1174 Deep Metazoan Phylogeny [HU 956/8]. IE acknowledges support by a grant of the Wiener Wissenschafts-, Forschungs- und Technologie Fonds (WWTF) to Arndt von Haezeler, and from the DFG priority program SPP 1174 Deep Metazoan Phylogeny [HA 1628/9]. We thank two anonymous reviewers and the editor for their helpful comments, and Kathleen Rankin for correction of the language.

References
1. World Health Organization: World malaria report 2009. Geneva, 2009.
2. Boyd MF: Historical review. In: Malariaology. Volume 2 Edited by: Boyd, MF. Philadelphia: WB. Saunders; 1949:3-25.
3. Escalante AA, Aypa FJ: Phylogeny of the malarial genus Plasmodium, derived from rRNA gene sequences. Proc Natl Acad Sci USA 1994, 91:11373-11377.
4. Ollomo B, Durand P, Prugnolle F, Douzeuy E, Amathau C, Nkoghe D, Leroy E, Renaud F: A new malaria agent in African hominids. PLoS Pathog 2009, 5:e1000446.
5. Rich SM, Leendertz FH, Xu G, LeBreton M, Djoko CF, Aminiak MN, Takang EE, Diffo JL, Pike BL, Rosenthal BM, Formenty P, Boesch C, Aypa FJ, Wolfe ND: The origin of malignant malaria. Proc Natl Acad Sci USA 2009, 106:14902-14907.
6. Duval J, Fournet M, Neninnet E, Rouset D, Sadeuah SA, Goodman SM, Andriaholinirina NV, Randriananakelyosa M, Paul RE, Robert V, Aypa FJ, Arney F: African apes as reservoirs of Plasmodium falciparum and the origin and diversification of the Laverania subgenus. Proc Natl Acad Sci USA 2010, 107:10561-10566.
7. Krief S, Escalante AA, Pacheco MA, Mugisha L, André C, Hallwax M, Fischer A, Krief JM, Kasenene JM, Cranfield M, Cornejo OE, Chavatte JM, Lin C, Letourenoux F, Gruner AC, McCutchan TF, Réinsia L, Sounou G: On the diversity of malaria parasites in African apes and the origin of Plasmodium falciparum from Bonobos. PLoS Pathog 2010, 6:e1000765.
8. Liu W, Li Y, Leam GH, Rudicill RS, Robertson JD, Ndjogu JB, Sanz CM, Morgan DB, Locatelli S, Gonder MK, Kranzusch PJ, Wahls PD, Delaporte E, Mpoudi-Ngole E, Georgiev AV, Muller MN, Shaw GM, Peeters M, Sharp PM, Rayner JC, Hahn BH: Origin of the human malaria parasite Plasmodium falciparum in gorillas. Nature 2010, 467:420-425.
9. Prugnolle F, Durand P, Neel C, Ollomo B, Aypa FJ, Aymathau C, Etienne L, Mpoudi-Ngole E, Nkoghe D, Leroy E, Delaporte E, Peeters M, Renaud F: African great apes are natural hosts of multiple related malaria species, including Plasmodium falciparum. Proc Natl Acad Sci USA 2010, 107:1456-1463.
10. Hagner SC, Misof B, Maier WA, Kampen H: Bayesian analysis of new and old malaria parasite DNA sequence data demonstrates the need for more phylogenetic signal to clarify the descent of Plasmodium falciparum. Parasitol Res 2007, 101:493-503.
11. Maier AG, Rug M, O’Neill MT, Brown M, Chakravorty S, Szestak T, Chesson J, Wu Y, Hughes K, Coppel RL, Newbold C, Beeson JG, Craig A, Crabb BS, Peeters M, Sharp PM, Rayner JC, Hahn BH: The origin of the human malaria parasite Plasmodium falciparum-infected human erythrocytes. Cell 2008, 134:48-61.
12. Baruch DI, Pasloes BS, Singh HB, Bi X, Ma XC, Feldman M, Taraschi TF, Howard RJ: Cloning the P. falciparum gene encoding PEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. Cell 1997, 92:77-87.
13. Maiti M, Good RT, Rug M, Knupfer E, Cowman AF: Targeting malaria virulence and remodeling proteins to the host erythrocyte. Science 2004, 306:1930-1933.
25. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Spielmann T, Gilberger TW, MacKenzie JJ, Gómez ND, Bhattacharjee S, Mann S, Haldar K, Goldberg DE, Cowman AF: An aspartyl protease directly targets malaria effector proteins to the host cell. Nature 2010, 463:627-631.

26. Russo J, Babitt S, Malaldehvan V, Butler T, Okommen A, Goldberg DE: Plasmpesin V licenses Plasmodium proteins for export into the host erythrocyte. Nature 2010, 463:632-636.

28. Carlton JM, Adams JH, Silva JC, Bidwell SJ, Lorenzi H, Caler E, Crabtree J, Pick http://www.biomedcentral.com/1471-2148/11/167.
46. Crooks GE, Hon G, Chandonia JM, Brenner SE: WebLogo: a sequence logo generator. Genome Res 2004, 14:1188-1190.
47. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 2004, 340:783-795.
48. Kuo CH, Wares JP, Kissinger JC: The Apicomplexan whole-genome phylogeny: an analysis of incongruence among gene trees. Mol Biol Evol 2008, 25:2689-2698.
49. Talavera G, Castresana J: Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 2007, 56:564-577.
50. Strimmer K, Rambaut A: Inferring confidence sets of possibly misspecified gene trees. Proc Biol Sci 2006, 269:137-142.
51. Shimodaira H: An approximately unbiased test of phylogenetic tree selection. Syst Biol 2002, 51:492-508.
52. Dávalos LM, Perkins SL: Saturation and base composition bias explain phylogenetic conflict in Plasmodium. Genomics 2008, 91:433-442.
53. Silva JC, Egan A, Friedman R, Munro JB, Carlton JM, Hughes AL: Genome sequences reveal divergence times of malaria parasite lineages. Parasitology 2010, 1:1-13.
54. Philippe H, Brinkmann H, Lavrov DV, Littlewood DT, Manuel M, Wörheide G, Baurain D: Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol 2011, 9:e1000602.
55. Waters AP, Higgins DG, McCutchan TF: Plasmodium falciparum appears to have arisen as a result of lateral transfer between avian and human hosts. Proc Natl Acad Sci USA 1991, 88:3140-3144.
56. Escalante AA, Ayala FJ: Evolutionary origin of Plasmodium and other Apicomplexa based on rRNA genes. Proc Natl Acad Sci USA 1995, 92:5793-5797.
57. Kissinger JC, Souza PC, Soares CT, Paul R, Wahl AM, Rathore D, McCutchan TF, Kretli AU: Molecular phylogenetic analysis of the avian malarial parasite Plasmodium (Novellya) juxtanucleare. J Parasitol 2002, 88:769-773.
58. Perkins SL, Schall JJ: A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences. J Parasitol 2002, 88:972-978.
59. Leclerc MC, Hugot JP, Durand P, Renaud F, Carlton JM, Hughes AL: Evolutionary relationships between 15 Plasmodium species from new and old world primates (including humans): an 18S rDNA cladistic analysis. Parasitology 2004, 129:677-684.
60. Roy SW, Trima M: Origins of human malaria: rare genomic changes and full mitochondrial genomes confirm the relationship of Plasmodium falciparum to other mammalian parasites but complicate the origins of Plasmodium vivax. Mol Biol Evol 2008, 25:1192-1198.
61. Martinsen ES, Perkins SL, Schall JJ: A three-genome phylogeny of malaria parasites (Plasmodium and closely related genera): evolution of life-history traits and host switches. Mol Phylogenet Evol 2008, 47:261-273.
62. Philippe H, Derelle R, Lopez P, Pick K, Borchelli C, Bouy-Esnault N, Vacelet J, Renaud E, Houliston E, Quéinnec E, Da Silva C, Windker P, Le Guyader H, Leys S, Jackson DJ, Schreiber F, Erpenbeck D, Morgenstern B, Wörheide G, Manuel M: Phylogenomics revives traditional views on deep animal relationships. Curr Biol 2009, 19:706-712.
63. Delluc F, Brinkmann H, Philippe H: Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 2005, 6:361-375.
64. Graham SW, Olmstead RG, Barrett SC: Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. Mol Biol Evol 2002, 19:1769-1781.
65. McGhee B: The adaptation of the avian malaria parasite Plasmodium lophurae to a continuous existence in infant mice. J Infect Dis 1951, 88:86-97.
66. Conant GC, Wolfe KH: Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 2008, 9:938-950.
67. Boddy JA, Montez RL, Simpson RJ, Cowman AF: Role of the Plasmodium export element in trafficking parasite proteins to the infected erythrocyte. Traffic 2009, 10:285-299.
68. Qari SH, Shi YP, Piensziszczuk N, Collins WE, Lal AA: Phylogenetic relationship among the malaria parasites based on small subunit rRNA gene sequences: monophyletic nature of the human malaria parasite, Plasmodium falciparum. Mol Phylogenet Evol 1996, 6:157-165.
69. Escalante AA, Freeland DE, Collins WE, Lal AA: The evolution of primate malaria parasites based on the gene encoding cytochrome b from the linear mitochondrial genome. Proc Natl Acad Sci USA 1998, 95:8124-8129.
70. Rathore D, Wahl AM, Sullivan M, McCutchan TF: A phylogenetic comparison of gene trees constructed from plastid, mitochondrial and genomic DNA of Plasmodium species. Mol Biochem Parasitol 2001, 114:89-94.
71. Huson DH: SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 1998, 14:68-73.