Integrative proteomics and bioinformatic prediction enable a high-confidence apicoplast proteome in malaria parasites

Michael J. Boucher[1,2], Sreejoyee Ghosh[2], Lichao Zhang[3], Avantika Lal[4,10], Se Won Jang[5,10], An Ju[6,11], Shuying Zhang[6,11], Xinzi Wang[6,11], Stuart A. Ralph[7], James Zou[8,9], Joshua E. Elias[3], and Ellen Yeh[1,2,4,9*]

[1]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305, United States of America

[2]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305, United States of America

[3]Department of Chemical and Systems Biology, Stanford University School of Medicine, Stanford, CA 94305, United States of America

[4]Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, United States of America

[5]Department of Computer Science, Stanford University, Stanford, CA 94305, United States of America

[6]Department of Bioengineering, Stanford University, Stanford, CA 94305, United States of America

[7]Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Parkville, Vic 3010, Australia

[8]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305, United States of America[9]Chan Zuckerberg Biohub, San Francisco, CA 94158, United States of America

[10]These authors contributed equally

[11]These authors contributed equally

*Correspondence: ellenyeh@stanford.edu

Keywords: malaria, apicoplast, BioID, proteome, neural network

Abstract

Malaria parasites (*Plasmodium* spp.) and related apicomplexan pathogens contain a non-photosynthetic plastid called the apicoplast. Derived from an unusual secondary eukaryote-eukaryote endosymbiosis, the apicoplast is a fascinating organelle whose function and biogenesis rely on a complex amalgamation of bacterial and algal pathways. Because these pathways are distinct from the human host, the apicoplast is an excellent source of novel antimalarial targets. Despite its biomedical importance and evolutionary significance, the absence of a reliable apicoplast proteome has limited most studies to the handful of pathways identified by homology to bacteria or primary chloroplasts, precluding our ability to study the most novel apicoplast pathways. Here we combine proximity biotinylation-based proteomics (BioID) and a new machine learning algorithm to generate a high-confidence apicoplast proteome consisting of 346 proteins. Critically, the high accuracy of this proteome significantly outperforms previous prediction-based methods and extends beyond other BioID studies of unique parasite compartments. Half of identified proteins have unknown function, and 77% are predicted to be important for normal blood-stage growth. We validate the apicoplast localization of a subset of novel proteins and show that an ATP-binding cassette protein ABCF1 is essential for blood-stage survival and plays a previously unknown role in apicoplast biogenesis. These findings indicate critical organellar functions for newly discovered apicoplast proteins. The apicoplast proteome will be an important resource for elucidating unique pathways derived from secondary endosymbiosis and prioritizing antimalarial drug targets.

Introduction

Identification of new antimalarial drug targets is urgently needed to address emerging resistance to all currently available therapies. However, nearly half of the *Plasmodium falciparum* genome encodes conserved proteins of unknown function [1], obscuring critical pathways required for malaria pathogenesis. The apicoplast is an essential, non-photosynthetic plastid found in *Plasmodium* spp. and related apicomplexan pathogens [2, 3]. This unusual organelle is an enriched source of both novel cellular pathways and parasite-specific drug targets [4]. It was acquired by secondary (i.e., eukaryote-eukaryote) endosymbiosis and has evolutionarily diverged from the primary endosymbiotic organelles found in model organisms. While some aspects of apicoplast biology are shared with bacteria, mitochondria, and primary chloroplasts, many are unique to the secondary plastid in this parasite lineage. For example, novel translocons import apicoplast proteins through specialized membranes derived from secondary endosymbiosis [5-8], while the parasite's pared-down metabolism necessitates export of key metabolites from the apicoplast using as-yet unidentified small molecule transporters [9, 10].

These novel cellular pathways, which are also distinct from human host cells, can be exploited for antimalarial drug discovery. Indeed, antimalarials that target apicoplast pathways are currently in use as prophylactics or partner drugs (doxycycline, clindamycin) or have been tested in clinical trials (fosmidomycin) [11-15]. However, known apicoplast drug targets have been limited to the handful of pathways identified by homology to plastid-localized pathways in model organisms. Meanwhile the number of druggable apicoplast targets, including those in unique secondary plastid pathways, is likely more extensive [16].

A major hurdle to identifying novel, parasite-specific pathways and prioritizing new apicoplast targets is the lack of a well-defined organellar proteome. So far, the apicoplast has not

been isolated in sufficient yield or purity for traditional organellar proteomics. Instead, large-scale, unbiased identification of apicoplast proteins has relied on bioinformatic prediction of apicoplast targeting sequences [17-19]. These prediction algorithms identify hundreds of putative apicoplast proteins but contain numerous false positives. Confirmation of these low-confidence candidate apicoplast proteins is slow due to the genetic intractability of *P. falciparum* parasites. Unbiased identification of apicoplast proteins in an accurate and high-throughput manner would significantly enhance our ability to study novel apicoplast pathways and validate new antimalarial drug targets.

BioID and other cellular proximity labeling methods are attractive techniques for identification of organellar proteins [20, 21]. In BioID, a promiscuous biotin ligase, BirA*, is fused to a bait protein and catalyzes biotinylation of neighbor proteins in intact cells. Proximity labeling methods have been used for unbiased proteomic profiling of subcellular compartments in diverse parasitic protozoa, including *Plasmodium* spp. [22-29]. Here we used BioID to perform large-scale identification of *P. falciparum* apicoplast proteins during asexual blood-stage growth. Extending beyond previous BioID studies of unique parasite compartments, we achieved high positive predictive value of true apicoplast proteins by implementing an endoplasmic reticulum (ER) negative control to remove frequent contaminants expected based on the trafficking route of apicoplast proteins. Furthermore, higher coverage was achieved by using the proteomic dataset to develop an improved neural network prediction algorithm, PlastNN. We now report a high-confidence apicoplast proteome of 346 proteins rich in novel and essential functions.

Results

The promiscuous biotin ligase BirA* is functional in the *P. falciparum* apicoplast and endoplasmic reticulum

To target the promiscuous biotin ligase BirA* to the apicoplast, the *N*-terminus of a GFP-BirA* fusion protein was modified with the apicoplast-targeting leader sequence from acyl carrier protein (ACP) (Fig 1A). Since apicoplast proteins transit the parasite ER en route to the apicoplast [30], we also generated a negative control in which GFP-BirA* was targeted to the ER via an *N*-terminal signal peptide and a *C*-terminal ER-retention motif (Fig 1A). Each of these constructs was integrated into an ectopic locus in Dd2^{attB} parasites [31] to generate BioID-Ap and BioID-ER parasites (S1A Fig). Immunofluorescence co-localization and live imaging of these parasites confirmed GFP-BirA* localization to either the apicoplast or the ER, respectively (Fig 1B and S1B Fig).

To test the functionality of the GFP-BirA* fusions in the apicoplast and ER, we labeled either untransfected Dd2^{attB}, BioID-Ap, or BioID-ER parasites with DMSO or 50 µM biotin and assessed biotinylation by western blotting and fixed-cell fluorescent imaging. As has been reported [28], significant labeling of GFP-BirA*-expressing parasites above background was achieved even in the absence of biotin supplementation, suggesting that the 0.8 µM biotin in RPMI growth medium is sufficient for labeling (Fig 1C). Addition of 50 µM biotin further increased protein biotinylation. Fluorescence imaging of biotinylated proteins revealed staining that co-localized with the respective apicoplast- or ER-targeted GFP-BirA* fusion proteins (Fig 1D). These results confirm that GFP-BirA* fusions are active in the *P. falciparum* apicoplast and ER and can be used for compartment-specific biotinylation of proteins.

Proximity-dependent labeling (BioID) generates an improved apicoplast proteome dataset

For large-scale identification of apicoplast proteins, biotinylated proteins from late-stage BioID-Ap and BioID-ER parasites were purified using streptavidin-conjugated beads and identified by mass spectrometry. A total of 728 unique *P. falciparum* proteins were detected in the apicoplast and/or ER based on presence in at least 2 of 4 biological replicates and at least 2 unique spectral matches in any single mass spectrometry run (Fig 2A and S1 Table). The abundance of each protein in apicoplast and ER samples was calculated by summing the total MS1 area of all matched peptides and normalizing to the total MS1 area of all detected *P. falciparum* peptides within each mass spectrometry run.

To assess the ability of our dataset to distinguish between true positives and negatives, we generated control lists of 96 known apicoplast and 451 signal peptide-containing non-apicoplast proteins based on published localizations and pathways (S2 Table). Consistent with an enrichment of apicoplast proteins in BioID-Ap samples, we observed a clear separation of known apicoplast and non-apicoplast proteins based on apicoplast:ER abundance ratio (Fig 2A). Using receiver operating characteristic (ROC) curve analysis (Fig 2B), we set a threshold of apicoplast:ER abundance ratio ≥5-fold for inclusion of 187 proteins in the BioID apicoplast proteome, which maximized sensitivity while minimizing false positives (Fig 2A, dotted line; S1 Table). This dataset included 50 of the 96 positive control proteins for a sensitivity of 52% (95% CI: 42-62%). None of the original 451 negative controls were present above the ≥5-fold enrichment threshold, but manual inspection of this list identified 5 likely false positives not present on our initial list (S1 Table) for a positive predictive value (PPV; the estimated fraction of proteins on the list that are true positives) of 91% (95% CI: 80-96%).

To benchmark our dataset against the current standard for large-scale identification of apicoplast proteins, we compared the apicoplast BioID proteome to the predicted apicoplast

144 proteomes from three published bioinformatic algorithms: PATS [17], PlasmoAP [18], and
145 ApicoAP [19] (S3 Table). At 52% sensitivity, apicoplast BioID identified fewer known
146 apicoplast proteins than PATS or PlasmoAP, which had sensitivities of 89% and 84%,
147 respectively, but outperformed the 40% sensitivity of ApicoAP (Fig 2C). All three algorithms as
148 well as apicoplast BioID achieved high negative predictive values (NPV), since NPV is
149 influenced by the larger number of true negatives (known non-apicoplast proteins) than true
150 positives (known apicoplast) from literature data (S2A Fig). We expected that the advantages of
151 apicoplast BioID would be its improved discrimination between true and false positives (Fig 2A)
152 and the ability to detect proteins without classical targeting presequences. Indeed, bioinformatic
153 algorithms had poor PPVs ranging from 19-36% compared to the 91% PPV of BioID (Fig 2D).
154 Even a dataset consisting only of proteins predicted by all three algorithms achieved a PPV of
155 just 25%. Similarly, the specificity of BioID outperformed that of the bioinformatic algorithms
156 (S2B Fig). Consistent with the low PPVs of the bioinformatic algorithms, many proteins
157 predicted by these algorithms are not enriched in BioID-Ap samples and are likely to be false
158 positives (S3 Fig). Altogether, identification of apicoplast proteins using BioID provided a
159 dramatic improvement in prediction performance over bioinformatic algorithms.
160
161 **Apicoplast BioID identifies proteins of diverse functions in multiple subcompartments**
162 To determine whether lumenally targeted GFP-BirA* exhibited any labeling preferences,
163 we assessed proteins identified based on the presence of transmembrane domains, their sub-
164 organellar localization, and their functions. First, we determined the proportion of the 187
165 proteins identified by apicoplast BioID that are membrane proteins. To ensure that proteins were
166 not classified as membrane proteins solely due to misclassification of a signal peptide as a

transmembrane domain, we considered a protein to be in a membrane only if it contained at least one predicted transmembrane domain more than 80 amino acids from the protein's *N*-terminus (as determined by annotation in PlasmoDB). These criteria suggested that 11% of identified proteins (20/187) were likely membrane proteins (Fig 3A), indicating that lumenal GFP-BirA* can label apicoplast membrane proteins.

Second, apicoplast proteins may localize to one or multiple sub-compartments defined by the four apicoplast membranes. It was unclear whether BirA* targeted to the lumen would label proteins in non-lumenal compartments. Based on literature descriptions, we classified the 96 known apicoplast proteins on our positive control list as either lumenal (present in lumenal space or on the innermost apicoplast membrane) or non-lumenal (all other sub-compartments) and determined the proportion that were identified in our dataset. Apicoplast BioID identified 56% (45/81) of the classified lumenal proteins and 33% (5/15) of the non-lumenal proteins (Fig 3B), suggesting that the GFP-BirA* bait used can label both lumenal and non-lumenal proteins but may have a preference for lumenal proteins (though this difference did not reach statistical significance).

Finally, we characterized the functions of proteins identified by apicoplast BioID. We grouped positive control apicoplast proteins into functional categories and assessed the proportion of proteins identified from each functional group (Fig 3C). BioID identified a substantial proportion (67-100%) of proteins in four apicoplast pathways that are essential in blood stage and localize to the apicoplast lumen, specifically DNA replication, protein translation, isoprenoid biosynthesis, and iron-sulfur cluster biosynthesis. Conversely, BioID identified few proteins involved in heme or fatty acid biosynthesis (0% and 17%, respectively), which are lumenal pathways that are non-essential in the blood-stage and which are likely to be

more abundant in other life cycle stages [32-36]. We achieved moderate coverage of proteins involved in protein quality control (44%) and redox regulation (38%). Consistent with the reduced labeling of non-lumenal apicoplast proteins, only a small subset (29%) of proteins involved in import of nuclear-encoded apicoplast proteins were identified. Overall, apicoplast BioID identified soluble and membrane proteins of diverse functions in multiple apicoplast compartments with higher coverage for lumenal proteins required during blood-stage infection.

The PlastNN algorithm expands the predicted apicoplast proteome with high accuracy

Apicoplast BioID provided the first experimental profile of the blood-stage apicoplast proteome but is potentially limited in sensitivity due to 1) difficulty in detecting low abundance peptides in complex mixtures; 2) inability of the promiscuous biotin ligase to access target proteins that are buried in membranes or protein complexes; or 3) stage-specific protein expression. Currently available bioinformatic predictions of apicoplast proteins circumvent these limitations, albeit at the expense of a low PPV (Fig 2D). We reasoned that increasing the number of high-confidence apicoplast proteins used to train algorithms could improve the accuracy of a prediction algorithm while maintaining high sensitivity. In addition, inclusion of exported proteins that traffic through the ER, which are common false positives in previous prediction algorithms, would also improve our negative training set.

We used our list of previously known apicoplast proteins (S2 Table) as well as newly-identified apicoplast proteins from BioID (S1 Table) to construct a positive training set of 205 apicoplast proteins (S4 Table). As a negative training set, we used our previous list of 451 signal peptide-containing non-apicoplast proteins (S2 Table). For each of the 656 proteins in the training set, we calculated the frequencies of all 20 canonical amino acids in a 50 amino acid

region immediately following the predicted signal peptide cleavage site. In addition, given that apicoplast proteins have a characteristic transcriptional profile in blood-stage parasites [37] and that analysis of transcriptional profile has previously enabled identification of apicoplast proteins in the related apicomplexan *Toxoplasma gondii* [38], we obtained transcript levels at 8 time points during intraerythrocytic development from previous RNA-Seq data [39]. Altogether, each protein was represented by a vector of dimension 28 (20 amino acid frequencies plus 8 transcript levels). These 28-dimensional vectors were used as inputs to train a neural network with 3 hidden layers (Fig 4A and S5 Table). Six-fold cross-validation was used for training, wherein the training set was divided into 6 equal parts (folds) to train 6 separate models. Each time, 5 folds were used to train the model and 1 fold to measure the performance of the trained model.

We named this model PlastNN (ApicoPLAST Neural Network). PlastNN recognized apicoplast proteins with a cross-validation accuracy of 96 ± 2% (mean ± SD across 6 models), along with sensitivity of 95 ± 5% and PPV of 95 ± 4% (Fig 4B). This performance was higher than logistic regression on the same dataset (average accuracy = 91%; S6 Table). Combining the transcriptome features and the amino acid frequencies improves performance: the same neural network architecture with amino acid frequencies alone as input resulted in a lower average accuracy of 91%, while using transcriptome data alone resulted in an average accuracy of 90% (S6 Table). Comparison of the performance of PlastNN to existing prediction algorithms indicates that PlastNN distinguishes apicoplast and non-apicoplast proteins with higher accuracy than any previous prediction method (Fig 4C). To identify new apicoplast proteins, PlastNN was used to predict the apicoplast status of 450 predicted signal peptide-containing proteins that were not in our positive or negative training sets. Since PlastNN is composed of 6 models, we designated proteins as "apicoplast" if plastid localization was predicted by ≥4 of the 6 models.

PlastNN predicts 118 out of the 450 proteins to be targeted to the apicoplast (S7 Table). Combining these results with those from apicoplast BioID (S1 Table) and with experimental localization of proteins from the literature (S2 Table) yielded a compiled proteome of 346 putative nuclear-encoded apicoplast proteins (S8 Table).

The apicoplast proteome contains novel and essential proteins

To determine whether candidate apicoplast proteins from this study have the potential to reveal unexplored parasite biology or are candidate antimalarial drug targets, we assessed the novelty and essentiality of the identified proteins. We found that substantial fractions of the BioID and PlastNN proteomes (49% and 71%, respectively) and 50% of the compiled apicoplast proteome represented proteins that could not be assigned to an established apicoplast pathway and therefore might be involved in novel organellar processes (Fig 5A). Furthermore, we identified orthologs of identified genes in the 150 genomes present in the OrthoMCL database [40]: 39% of the compiled apicoplast proteome were unique to apicomplexan parasites, with 58% of these proteins found only in *Plasmodium* spp. (Fig 5B). Of the 61% of proteins that were conserved outside of the Apicomplexa, we note that many of these contain conserved domains or are components of well-established pathways, such as DNA replication, translation, and metabolic pathways (S8 Table). This analysis indicates that many of the newly identified proteins are significantly divergent from proteins in their metazoan hosts.

Consistent with the critical role of the apicoplast in parasite biology, a genome-scale functional analysis of genes in the rodent malaria parasite *P. berghei* showed that numerous apicoplast proteins are essential for blood-stage survival [41]. Using this dataset, we found that 77% of proteins in the compiled apicoplast proteome that had *P. berghei* homologs analyzed by

PlasmoGEM were important for normal blood-stage parasite growth (Fig 5C). Notably, of 49 proteins that were annotated explicitly with "unknown function" in their gene description and for which essentiality data are available, 38 are important for normal parasite growth, indicating that the high rate of essentiality for apicoplast proteins is true of both previously known and newly discovered proteins. In concordance with the PlasmoGEM data, recent genome-scale transposon mutagenesis in *P. falciparum* [42] identified 75% of proteins in the compiled apicoplast proteome as non-mutable (Fig 5D), suggesting essential functions in the blood stage. Overall, these data suggest that we have identified dozens of novel proteins that are likely critical for apicoplast biology.

Localization of candidate apicoplast proteins identifies novel proteins of biological interest

Our analyses of the apicoplast BioID and PlastNN datasets suggested that these approaches enabled accurate, large-scale identification of apicoplast proteins (Figs 2D and 4C) and included many proteins of potential biological interest due to their novelty or their essentiality in the blood stage (Fig 5). As proof-of-concept of the utility of these datasets, several newly identified apicoplast proteins were experimentally validated. Fortuitously, while this manuscript was in preparation, 7 new apicoplast membrane proteins in *P. berghei* were validated by Sayers et al. [43]. Of these, apicoplast BioID identified the *P. falciparum* homologs of 3 proteins (PF3D7_1145500/ABCB3, PF3D7_0302600/ABCB4, and PF3D7_1021300) and PlastNN identified one (PF3D7_0908100). In addition to these, we also selected 4 candidates from apicoplast BioID and 2 from PlastNN to validate.

From the BioID list (S1 Table), we chose a rhomboid protease homolog ROM7 (PF3D7_1358300) and 3 conserved *Plasmodium* proteins of unknown function (PF3D7_0521400, PF3D7_1472800, and PF3D7_0721100) and generated cell lines expressing

283 *C*-terminal GFP fusions from an ectopic locus in Dd2^{attB} parasites. With the exception of ROM7,
284 which was chosen because of the biological interest of rhomboid proteases, we focused on
285 proteins of unknown function to begin characterizing the large number of unannotated proteins
286 in the *Plasmodium* genome (see Materials and Methods for additional candidate selection
287 criteria).

288 To assess the apicoplast localization of each candidate, we first detected the apicoplast-
289 dependent cleavage of each protein as a marker of its import. Most nuclear-encoded apicoplast
290 proteins are proteolytically processed to remove *N*-terminal targeting sequences following
291 successful import into the apicoplast [44, 45]. This processing is abolished in parasites rendered
292 "apicoplast-minus" by treatment with an inhibitor (actinonin) to cause apicoplast loss [16, 46].
293 Comparison of protein molecular weight in apicoplast-intact and -minus parasites showed that
294 ROM7, PF3D7_1472800, and PF3D7_0521400 (but not PF3D7_0721100) were cleaved in an
295 apicoplast-dependent manner (Fig 6A).

296 Next, we demonstrated co-localization of these three proteins with the apicoplast marker
297 ACP by co-immunofluorescence analysis (co-IFAs; Fig 6B). ROM7, PF3D7_1472800, and
298 PF3D7_0521400 clearly co-localized with ACP. PF3D7_0721100 localized to few large puncta
299 not previously described for any apicoplast protein, which partly co-localized with the apicoplast
300 marker ACP (Fig 6B and S4 Fig) but also appeared adjacent to ACP staining (Fig 6B and S4 Fig,
301 arrowheads).

302 Finally, we localized the candidate-GFP fusions by live fluorescence microscopy and
303 assessed the apicoplast dependence of their localization. ROM7-GFP, PF3D7_1472800-GFP,
304 and PF3D7_0521400-GFP localized to branched structures characteristic of the apicoplast (S5
305 Fig). Upon actinonin treatment to render parasites "apicoplast-minus," these proteins

306 mislocalized to diffuse puncta (S5 Fig) previously observed for known apicoplast proteins [46].
307 Interestingly, while in untreated live parasites PF3D7_0721100-GFP again localized to a few
308 large bright puncta, this protein also relocalized to the typical numerous diffuse puncta seen for
309 genuine apicoplast proteins in apicoplast-minus parasites (S5 Fig).
310 Taken together, these data validate the apicoplast localization of ROM7,
311 PF3D7_0521400, and PF3D7_1472800. Though transit peptide cleavage and the characteristic
312 branched structure were not detected for PF3D7_0721100, partial co-localization with ACP and
313 the mislocalization of PF3D7_0721100-GFP to puncta characteristic of apicoplast-minus
314 parasites indicates that this protein may also be a true apicoplast protein. Further studies using
315 either endogenously tagged protein or antibody raised against endogenous protein will be
316 necessary to better characterize this localization.
317 From the PlastNN list (S7 Table), we selected 2 proteins of unknown function,
318 PF3D7_1349900 and PF3D7_1330100. As above, each protein was appended with a *C*-terminal
319 GFP tag and expressed as a second copy in Dd2^{attB} parasites. In agreement with apicoplast
320 localization for each of these proteins, actinonin-mediated apicoplast loss caused loss of transit
321 peptide processing (Fig 7A) and redistribution from a branched structure to diffuse puncta (S6
322 Fig). Furthermore, both proteins co-localized with the apicoplast marker ACP (Fig 7B).
323 Altogether, we confirmed the apicoplast localization of 5 novel apicoplast proteins, with
324 a sixth protein (PF3D7_0721100) having potential apicoplast localization. These results,
325 combined with validation of 4 apicoplast membrane proteins predicted in our datasets by Sayers
326 et al., show that the apicoplast BioID and PlastNN datasets can successfully be used to prioritize
327 apicoplast proteins of biological interest.
328

A novel apicoplast protein ABCF1 is essential and required for organelle biogenesis

Given the potential of ATP binding cassette (ABC) proteins as drug targets, we sought to experimentally validate the essentiality of newly discovered apicoplast ABC proteins and assess their roles in metabolism or organelle biogenesis. Apicoplast BioID identified four ABC proteins: 3 ABCB-family proteins (ABCB3, ABCB4, and ABCB7) and an ABCF-family protein (ABCF1). We expected that these proteins might be important for apicoplast biology, as ABCB-family proteins are integral membrane proteins that typically act as small molecule transporters and ABCF-family proteins, which do not contain transmembrane domains, are typically involved in translation regulation [47, 48]. We pursued reverse genetic characterization of ABCB7 (PF3D7_1209900) and ABCF1 (PF3D7_0813700), as the essentiality of ABCB3 and ABCB4 has been previously studied [43, 49].

To assess localization and function of ABCB7 and ABCF1, we modified their endogenous loci to contain a *C*-terminal triple HA tag and tandem copies of a tetracycline repressor (TetR)-binding RNA aptamer in the 3′ UTR of either gene (S7 Fig) [50, 51]. Co-IFA confirmed ABCF1-3xHA colocalization with the apicoplast marker ACP (Fig 8A). ABCB7-3xHA localized to elongated structures that may be indicative of an intracellular organelle but rarely co-localized with ACP, indicating that it has a primarily non-apicoplast localization and is likely a false positive from the BioID dataset (S8A Fig).

Taking advantage of the TetR-binding aptamers in the 3′ UTR of ABCF1, we determined the essentiality and knockdown phenotype of this protein. In the presence of anhydrotetracycline (ATc), binding of the aptamer by a TetR-DOZI repressor is inhibited and ABCF1 is expressed. Upon removal of ATc, repressor binding blocks gene expression [50, 51]. Knockdown of ABCF1 caused robust parasite growth inhibition (Fig 8B-C). Growth inhibition of ABCF1-

352 deficient parasites was reversed in the presence of isopentenyl pyrophosphate (IPP) (Fig 8C),
353 which bypasses the need for a functional apicoplast [46], indicating that ABCF1 has an essential
354 apicoplast function. Essential apicoplast functions can be placed into two broad categories: those
355 involved in organelle biogenesis, and those involved solely in IPP production. Disruption of
356 proteins required for organelle biogenesis causes apicoplast loss, while disruption of proteins
357 involved in IPP production does not [16, 46, 52]. We determined whether knockdown of ABCF1
358 caused apicoplast loss by assessing 1) absence of the apicoplast genome, 2) loss of transit peptide
359 processing of nuclear-encoded apicoplast proteins, and 3) relocalization of apicoplast proteins to
360 puncta. Indeed, the apicoplast:nuclear genome ratio drastically decreased in ABCF1 knockdown
361 parasites beginning 1 cycle after knockdown (Fig 8D), and western blot showed that the
362 apicoplast protein ClpP was not processed in ABCF1 knockdown parasites (Fig 8E).
363 Furthermore, IFA of the apicoplast marker ACP confirmed redistribution from an intact plastid
364 to diffuse cytosolic puncta (Fig 8F). In contrast to ABCF1, a similar knockdown of ABCB7
365 caused no observable growth defect after four growth cycles despite significant reduction in
366 protein levels (S8B-C Fig). Together, these results show that ABCF1 is a novel and essential
367 apicoplast protein with a previously unknown function in organelle biogenesis.
368
369 **Discussion**
370 Since the discovery of the apicoplast, identification of its proteome has been a pressing
371 priority. We report the first large-scale proteomic analysis of the apicoplast in blood-stage
372 malaria parasites, which identified 187 candidate proteins with 52% sensitivity and 91% PPV. A
373 number of groups have also profiled parasite-specific membrane compartments using proximity
374 biotinylation but observed contamination with proteins in or trafficking through the ER,

preventing accurate identification of these proteomes without substantial manual curation and validation [23, 24, 26-29]. This background labeling is expected since proteins traffic through the ER en route to several parasite-specific compartments, including the parasitophorous vacuole, host cytoplasm, food vacuole, and invasion organelles. The high specificity of our apicoplast BioID proteome depended on 1) the use of a control cell line expressing ER-localized GFP-BirA* to detect enrichment of apicoplast proteins from background ER labeling and 2) strong positive and negative controls to set an accurate threshold. We suspect a similar strategy to detect nonspecific ER background may also improve the specificity of proteomic datasets for other parasite-specific, endomembrane-derived compartments.

Leveraging our successful proteomic analysis, we used these empirical data as an updated training set to also improve computational predictions of apicoplast proteins. PlastNN identified an additional 118 proteins with 95% sensitivity and 95% PPV. Although two previous prediction algorithms, PATS and ApicoAP, also applied machine learning to the problem of transit peptide prediction, we reasoned that their low accuracy arose from the small training sets used (ApicoAP) and the use of cytosolic as well as endomembrane proteins in the negative training set (PATS). By using an expanded positive training set based on proteomic data and limiting our training sets to only signal peptide-containing proteins, we developed an algorithm with higher sensitivity than BioID and higher accuracy than previous apicoplast protein prediction models. Inevitably some false positives from the BioID dataset would have been used for neural network training and cross-validation. While this may slightly influence the PPV of the PlastNN list, we expect that the substantially larger fraction of true positives in the training set mitigated the effects of any false positives. Importantly, as more apicoplast and non-apicoplast proteins in *P. falciparum* parasites are experimentally validated, updated training sets can be used to re-train

PlastNN. Moreover, PlastNN suggests testable hypotheses regarding the contribution of sequence-based and temporal regulation to protein trafficking in the ER.

Overall, we have compiled a high-confidence apicoplast proteome of 346 proteins that are rich in novel and essential functions (Fig 5). This proteome likely represents a majority of soluble apicoplast proteins, since 1) our bait for proximity biotinylation targeted to the lumen and 2) most soluble proteins use canonical targeting sequences that can be predicted. An important next step will be to expand the coverage of apicoplast membrane proteins, which more often traffic via distinctive routes [53, 54]. Performing proximity biotinylation with additional bait proteins may identify such atypical apicoplast proteins. In the current study, our bait was an inert fluorescent protein targeted to the apicoplast lumen to minimize potential toxicity of the construct. The success of this apicoplast GFP bait gives us confidence to attempt more challenging baits, including proteins localized to sub-organellar membrane compartments or components of the protein import machinery. Performing apicoplast BioID in liver and mosquito stages may also define apicoplast functions in these stages. This compiled proteome represents a substantial improvement upon previous bioinformatics predictions of apicoplast proteins and provides a strong foundation for further refinement. In analogy to progress on the mammalian mitochondrial proteome, which over the course of decades has been expanded and refined by a combination of proteomic, computational, and candidate-based approaches [55, 56], we expect that future proteomic, computational, and candidate-based approaches to identify apicoplast proteins will be critical for ultimately determining a comprehensive apicoplast proteome,

Organellar proteomes are valuable hypothesis-generating tools. Already several candidates of biological interest based on their biochemical function annotations were validated. We demonstrated an unexpected role for the ATP-binding cassette protein *Pf*ABCF1 in

apicoplast biogenesis. ABCF proteins are understudied compared to other ABC-containing proteins but tend to have roles in translation regulation [47]. An *E. coli* homolog, EttA, regulates translation initiation in response to cellular ATP levels [57, 58], and mammalian and yeast ABCF1 homologs also interact with ribosomes and regulate translation [59-62]. By analogy, *Pf*ABCF1 may regulate the prokaryotic translation machinery in the apicoplast, although the mechanistic basis for the severe defect in parasite replication upon loss of *Pf*ABCF1 is unclear.

We also validated *Pf*ROM7 as an apicoplast-localized rhomboid protease. Rhomboid proteases are a diverse family of intramembrane serine proteases found in all domains of life. In the Apicomplexa, rhomboids have been studied primarily for their roles in processing adhesins on the parasite cell surface [63], although the functions of most apicomplexan rhomboids are still unknown. Little is known about ROM7 other than that it appears to be absent from coccidians and was refractory to deletion in *P. berghei* [64, 65]. However, a rhomboid protease was recently identified as a component of symbiont-derived ERAD-like machinery (SELMA) that transports proteins across a novel secondary plastid membrane in diatoms [66], indicating that ROM7 may similarly play a role in apicoplast protein import in *Plasmodium* parasites. Neither *Pf*ABCF1 nor *Pf*ROM7 had known roles in the apicoplast prior to their identification in this study, underscoring the utility of unbiased approaches to identify new organellar proteins. Moreover, the apicoplast is one of few models for complex plastids that permits functional analysis of identified proteins to investigate the molecular mechanisms underpinning serial endosymbiosis. A summary of all candidate proteins validated in this study is shown in S9 Table.

A recent study aimed at identifying apicoplast membrane transporters highlights the difficulty in identifying novel apicoplast functions in the absence of a high-confidence proteome [43]. Taking advantage of the tractable genetics in murine *Plasmodium* species, Sayers et al.

444 screened 27 candidates in *P. berghei* for essentiality and apicoplast localization. Following >50
445 transfections, 3 essential and 4 non-essential apicoplast membrane proteins were identified. One
446 newly identified essential apicoplast membrane protein was then validated to be required for
447 apicoplast biogenesis in *P. falciparum*. In contrast, even though our study was not optimized to
448 identify membrane proteins, the combination of BioID and PlastNN identified 2 known
449 apicoplast transporters, 4 of the new apicoplast membrane protein homologs, and 56 additional
450 proteins predicted to contain at least one transmembrane domain. A focused screen of higher
451 quality candidates in *P. falciparum* is likely to be more rapid and yield the most relevant biology.
452 Our high-confidence apicoplast proteome will streamline these labor-intensive screens, focusing
453 on strong candidates for downstream biological function elucidation. As methods for analyzing
454 gene function in *P. falciparum* parasites continue to improve, this resource will become
455 increasingly valuable for characterizing unknown organellar pathways.
456

457 **Materials and Methods**

458 **Parasite growth**

459 *Plasmodium falciparum* Dd2^{attB} [31] (MRA-843) were obtained from MR4. NF54$^{Cas9+T7\ Polymerase}$
460 parasites [67] were a gift from Jacquin Niles. Parasites were grown in human erythrocytes (2%
461 hematocrit) obtained from the Stanford Blood Center in RPMI 1640 media (Gibco)
462 supplemented with 0.25% Albumax II (Gibco), 2 g/L sodium bicarbonate, 0.1 mM hypoxanthine
463 (Sigma), 25 mM HEPES, pH 7.4 (Sigma), and 50 μg/L gentamicin (Gold Biotechnology) at
464 37°C, 5% O_2, and 5% CO_2.
465

466 **Vector construction**

Oligonucleotides were purchased from the Stanford Protein and Nucleic Acid facility or IDT. gBlocks were ordered from IDT. Molecular cloning was performed using In-Fusion cloning (Clontech) or Gibson Assembly (NEB). Primer and gBlock sequences are available in S10 Table.

To generate the plasmid pRL2-ACP$_L$-GFP for targeting transgenes to the apicoplast, the first 55 amino acids from ACP were PCR amplified with primers MB015 and MB016 and were inserted in front of the GFP in the pRL2 backbone [68] via the AvrII/BsiWI sites. To generate pRL2-ACP$_L$-GFP-BirA* for targeting a GFP-BirA* fusion to the apicoplast, GFP was amplified from pLN-ENR-GFP using primers MB087 and MB088 and BirA* was amplified from pcDNA3.1 mycBioID (Addgene 35700) [20] using primers MB089 and MB090. These inserts were simultaneously cloned into BsiWI/AflII-digested pRL2-ACP$_L$-GFP to generate pRL2-ACP$_L$-GFP-BirA*. To generate pRL2-SP-GFP-BirA*-SDEL for targeting GFP-BirA* to the ER, SP-GFP-BirA*-SDEL was PCR amplified from pRL2-ACP$_L$-GFP-BirA* using primers MB093 and MB094 and was cloned into AvrII/AflII-digested pRL2-ACP$_L$-GFP. For GFP-tagging to confirm localization of proteins identified by apicoplast BioID, full-length genes were amplified from parasite cDNA with primers as described in S10 Table and were cloned into the AvrII/BsiWI sites of pRL2-ACP$_L$-GFP.

For CRISPR-Cas9-based editing of endogenous ABCB7 and ABCF1 loci, sgRNAs were designed using the eukaryotic CRISPR guide RNA/DNA design tool (http://grna.ctegd.uga.edu/). To generate a linear plasmid for CRISPR-Cas9-based editing, left homology regions were amplified with primers MB256 and MB257 (ABCB7) or MB260 and MB261 (ABCF1) and right homology regions were amplified with MB258 and MB259 (ABCB7) or MB262 and MB263 (ABCF1). For each gene, a gBlock containing the recoded coding sequence *C*-terminal of the CRISPR cut site and a triple HA tag was synthesized with appropriate overhangs for Gibson

Assembly. This fragment and the appropriate left homology region were simultaneously cloned into the FseI/ApaI sites of the linear plasmid pSN054-V5. Next, the appropriate right homology region and a gBlock containing the sgRNA expression cassette were simultaneously cloned into the AscI/I-SceI sites of the resultant vectors to generate the plasmids pSN054-ABCB7-TetR-DOZI and pSN054-ABCF1-TetR-DOZI.

Parasite transfection

Transfections were carried out using variations on the spontaneous uptake method [69, 70]. In the first variation, 100 µg of each plasmid was ethanol precipitated and resuspended in 30 µL sterile TE buffer and was added to 150 µL packed RBCs resuspended to a final volume of 400 µL in cytomix. The mixture was transferred to a 0.2 cm electroporation cuvette (Bio-Rad) and was electroporated at 310 V, 950 µF, infinity resistance in a Gene Pulser Xcell electroporation system (Bio-Rad) before allowing parasites to invade. Drug selection was initiated 3 days after transfection. Alternatively, 50 µg of each plasmid was ethanol precipitated and resuspended in 0.2 cm electroporation cuvettes in 100 µL TE buffer, 100 µL RPMI containing 10 mM HEPES-NaOH, pH 7.4, and 200 µL packed uninfected RBCs. RBCs were pulsed with 8 square wave pulses of 365 V x 1 ms separated by 0.1 s. RBCs were allowed to reseal for 1 hour in a 37°C water bath before allowing parasites to invade. Drug selection was initiated 4 days after transfection. All transfectants were selected with 2.5 µg/mL Blasticidin S (Research Products International). Additionally, BioID-ER parasites were selected with 125 µg/mL G418 sulfate (Corning) and ABCB7 and ABCF1 TetR-DOZI parasites were grown in the presence of 500 nM ATc. Transfections for generating BioID constructs (Fig 1) and expression of GFP-tagged candidates (Figs 6 and 7) were performed in the Dd2^{attB} background. Transfections for CRISPR

513 editing were performed with the NF54$^{Cas9+T7\ Polymerase}$ background and clonal parasite lines were
514 obtained by limiting dilution.

515 Correct modification of transfectant genomes was confirmed by PCR. Briefly, 200 µL of
516 2% hematocrit culture was pelleted and resuspended in water, and 2 µL of the resulting lysate
517 was used as template for PCR with Phusion polymerase (NEB). PCR targets and their
518 corresponding primer pairs are as follows: integrated *attL* site, p1 + p2; integrated *attR* site,
519 MW001 + MW003; unintegrated *attB* site, MW004 + MW003; ABCB7 unintegrated left
520 homology region (LHR), MB269 + MB270; ABCB7 integrated LHR, MB269 + MB255;
521 ABCB7 unintegrated right homology region (RHR), MB281 + MB278; ABCB7 integrated RHR,
522 MB276 + MB278; ABCF1 unintegrated LHR, MB271 + MB272; ABCF1 integrated LHR,
523 MB271 + MB255; ABCF1 unintegrated RHR, MB282 + MB283; ABCF1 integrated RHR,
524 MB276 + MB283.

525

526 **Biotin labeling**

527 To label parasites for analysis by streptavidin blot, fixed imaging, or mass spectrometry, cultures
528 of majority ring-stage parasites were treated with 50 µM biotin or with a DMSO vehicle-only
529 control. Cultures were harvested for analysis 16 hours later as majority trophozoites and
530 schizonts.

531

532 **Actinonin treatment and IPP rescue**

533 To generate apicoplast-minus parasites, ring-stage cultures were treated with 10 µM actinonin
534 (Sigma) and 200 µM IPP (Isoprenoids, LLC) and cultured for 3 days before analysis.

535

Western blotting

Parasites were separated from RBCs by lysis in 0.1% saponin and were washed in PBS. Parasite pellets were resuspended in PBS containing 1X NuPAGE LDS sample buffer with 50 mM DTT and were boiled at 95°C for 10 minutes before separation on NuPAGE or Bolt Bis-Tris gels and transfer to nitrocellulose. Membranes were blocked in 0.1% Hammarsten casein (Affymetrix) in 0.2X PBS with 0.01% sodium azide. Antibody incubations were performed in a 1:1 mixture of blocking buffer and TBST (Tris-buffered saline with Tween-20; 10 mM Tris, pH 8.0, 150 mM NaCl, 0.25 mM EDTA, 0.05% Tween 20). Blots were incubated with primary antibody for either 1 hour at room temperature or at 4°C overnight at the following dilutions: 1:20,000 mouse-α-GFP JL-8 (Clontech 632381); 1:20,000 rabbit-α-*Plasmodium* aldolase (Abcam ab207494); 1:1000 rat-α-HA 3F10 (Sigma 11867423001); 1:4000 rabbit-α-*Pf*ClpP [71]. Blots were washed once in TBST and were incubated for 1 hour at room temperature in a 1:10,000 dilution of the appropriate secondary antibody: IRDye 800CW donkey-α-rabbit; IRDye 680LT goat-α-mouse; IRDye 680LT goat-α-rat (LI-COR Biosciences). For detection of biotinylated proteins, blots were incubated with 1:1000 IRDye 680RD streptavidin for one hour at room temperature. Blots were washed three times in TBST and once in PBS before imaging on a LI-COR Odyssey imager.

Microscopy

For live imaging, parasites were settled onto glass-bottom microwell dishes (MatTek P35G-1.5-14-C) or Lab-Tek II chambered coverglass (ThermoFisher 155409) in PBS containing 0.4% glucose and 2 µg/mL Hoechst 33342 stain (ThermoFisher H3570).

For fixed imaging of biotinylated proteins in cells, biotin-labeled parasites were processed as in Tonkin et al. [72] with modifications. Briefly, parasites were washed in PBS and were fixed in 4% paraformaldehyde (Electron Microscopy Science 15710) and 0.015% glutaraldehyde (Electron Microscopy Sciences 16019) in PBS for 30 minutes. Cells were washed once in PBS, resuspended in PBS, and allowed to settle onto poly-L-lysine-coated coverslips (Corning) for 60 minutes. Coverslips were then washed once with PBS, permeabilized in 0.1% Triton X-100 in PBS for 10 minutes, and washed twice more in PBS. Cells were treated with 0.1 mg/mL sodium borohydride in PBS for 10 minutes, washed once in PBS, and blocked in 3% BSA in PBS. To visualize biotin-labeled proteins, coverslips were incubated with 1:1000 AlexaFluor 546-conjugated streptavidin (ThermoFisher S11225) for one hour followed by three washes in PBS. No labeling of GFP was necessary, as these fixation conditions preserve intrinsic GFP fluorescence [72]. Coverslips were mounted onto slides with ProLong Gold antifade reagent with DAPI (ThermoFisher) and were sealed with nail polish prior to imaging.

For immunofluorescence analysis, parasites were processed as above except that fixation was performed with 4% paraformaldehyde and 0.0075% glutaraldehyde in PBS for 20 minutes and blocking was performed with 5% BSA in PBS. Following blocking, primary antibodies were used in 5% BSA in PBS at the following concentrations: 1:500 rabbit-α-*Pf*ACP [73]; 1:1000 rabbit-α-*Pf*Bip1:1000 (a gift from Sebastian Mikolajczak and Stefan Kappe); 1:500 mouse-α-GFP JL-8 (Clontech 632381); 1:100 rat-α-HA 3F10 (Sigma 11867423001). Coverslips were washed three times in PBS, incubated with goat-α-rat 488 (ThermoFisher A-11006), goat-α-mouse 488 (ThermoFisher A11029), or donkey-α-rabbit 568 (ThermoFisher A10042) secondary antibodies at 1:3000, and washed three times in PBS prior to mounting as above.

Live and fixed cells were imaged with 100X, 1.4 NA or 100X, 1.35 NA objectives on an Olympus IX70 microscope with a DeltaVision system (Applied Precision) controlled with SoftWorx version 4.1.0 and equipped with a CoolSnap-HQ CCD camera (Photometrics). Images were taken in a single z-plane, with the exception of those presented in Figs 1D, 8A, and S8A, which were captured as a series of z-stacks separated by 0.2 µm intervals, deconvolved, and displayed as maximum intensity projections. Brightness and contrast were adjusted in Fiji (ImageJ) for display purposes. Image capture and processing conditions were identical for micrographs of the same cell line when multiple examples are displayed (S4 Fig) or when comparing untreated to actinonin-treated cells (S5 and S6 Figs).

Biotin pulldowns, mass spectrometry, and data analysis

Biotin-labeled parasites were harvested by centrifugation and were released from the host RBC by treatment with 0.1% saponin/PBS. Parasites were washed twice more with 0.1% saponin/PBS followed by PBS and were either used immediately for analysis or were stored at -80°C. Parasite pellets were resuspended in RIPA buffer [50 mM Tris-HCl, pH 7.4, 150 mM NaCl, 0.1% SDS, 0.5% sodium deoxycholate, 1% Triton X-100, 1 mM EDTA] containing a protease inhibitor cocktail (Pierce) and were lysed on ice for 30 minutes with occasional pipetting. Insoluble debris was removed by centrifugation at 16,000 xg for 15 minutes at 4°C. Biotinylated proteins were captured using High Capacity Streptavidin Agarose beads (Pierce) for 2 hours at room temperature. Beads were then washed three times with RIPA buffer, three times with SDS wash buffer [50 mM Tris-HCl, pH 7.4, 150 mM NaCl, 2% SDS], six times with urea wash buffer [50 mM Tris-HCl, pH 7.4, 150 mM NaCl, 8 M urea], and three times with 100 mM ammonium bicarbonate. Proteins were reduced with 5 mM DTT for 60 minutes at 37°C followed by

603 treatment with 14 mM iodoacetamide (Pierce) at room temperature for 45 minutes. Beads were
604 washed once with 100 mM ammonium bicarbonate and were digested with 10 µg/mL trypsin
605 (Promega) at 37°C overnight. The following day, samples were digested with an additional 5
606 µg/mL trypsin for 3-4 hours. Digested peptides were separated from beads by addition of either
607 35% or 50% final concentration acetonitrile, and peptides were dried on a SpeedVac prior to
608 desalting with C18 stage tips.

609 Desalted peptides were resuspended in 0.1% formic acid and analyzed by online capillary
610 nanoLC-MS/MS. Samples were separated on an in-house made 20 cm reversed phase column
611 (100 µm inner diameter, packed with ReproSil-Pur C18-AQ 3.0 µm resin (Dr. Maisch GmbH))
612 equipped with a laser-pulled nanoelectrospray emitter tip. Peptides were eluted at a flow rate of
613 400 nL/min using a two-step linear gradient including 2-25% buffer B in 70 min and 25-40% B
614 in 20 min (buffer A: 0.2% formic acid and 5% DMSO in water; buffer B: 0.2% formic acid and
615 5% DMSO in acetonitrile) in an Eksigent ekspert nanoLC-425 system (AB Sciex). Peptides were
616 then analyzed using a LTQ Orbitrap Elite mass spectrometer (Thermo Scientific). Data
617 acquisition was executed in data dependent mode with full MS scans acquired in the Orbitrap
618 mass analyzer with a resolution of 60000 and m/z scan range of 340-1600. The top 20 most
619 abundant ions with intensity threshold above 500 counts and charge states 2 and above were
620 selected for fragmentation using collision- induced dissociation (CID) with isolation window of
621 2 m/z, normalized collision energy of 35%, activation Q of 0.25 and activation time of 5 ms. The
622 CID fragments were analyzed in the ion trap with rapid scan rate. In additional runs, the top 10
623 most abundant ions with intensity threshold above 500 counts and charge states 2 and above
624 were selected for fragmentation using higher energy collisional dissociation (HCD) with
625 isolation window of 2 m/z, normalized collision energy of 35%, and activation time of 25 ms.

The HCD fragments were analyzed in the Orbitrap with a resolution of 15000. Dynamic exclusion was enabled with repeat count of 1 and exclusion duration of 30 s. The AGC target was set to 1000000, 50000, and 5000 for full FTMS scans, FTMSn scans and ITMSn scans, respectively. The maximum injection time was set to 250 ms, 250 ms, and 100 ms for full FTMS scans, FTMSn scans and ITMSn scans, respectively.

The resulting spectra were searched against a "target-decoy" sequence database [74] consisting of the PlasmoDB protein database (release 32, released April 19, 2017), the Uniprot human database (released February 2, 2015), and the corresponding reversed sequences using the SEQUEST algorithm (version 28, revision 12). The parent mass tolerance was set to 50 ppm and the fragment mass tolerance to 0.6 Da for CID scans, 0.02 Da for HCD scans. Enzyme specificity was set to trypsin. Oxidation of methionines was set as variable modification and carbamidomethylation of cysteines was set as static modification. Peptide identifications were filtered to a 1% peptide false discovery rate using a linear discriminator analysis [75]. Precursor peak areas were calculated for protein quantification.

Apicoplast protein prediction algorithms and positive/negative control apicoplast proteins

To generate updated lists of PATS-predicted apicoplast proteins, nuclear-encoded *P. falciparum* 3D7 proteins (excluding pseudogenes) from PlasmoDB version 28 (released March 30, 2016) were used to check for existence of a putative bipartite apicoplast targeting presequence using the artificial neural network predictor PATS [17].

Updated PlasmoAP-predicted apicoplast proteins were identified using the PlasmoDB version 32 proteome (released April 19, 2017) by first checking for the presequence of a predicted signal peptide using the neural network version of SignalP version 3.0 [76], and were

649 considered positive if they had a *D*-score above the default cutoff. The SignalP *C*-score was used
650 to predict the signal peptide cleavage position, and the remaining portion of the protein was
651 inspected for presence of a putative apicoplast transit peptide using the rules described for
652 PlasmoAP [18], implemented in a Perl script.

653 *P. falciparum* proteins predicted to localize to the apicoplast by ApicoAP were accessed
654 from the original paper [19]. Genes predicted to encode pseudogenes were excluded.

655 A positive control list of 96 high-confidence apicoplast proteins (S2 Table) was generated
656 based on either (1) published localization of that protein in *Plasmodium* parasites or *Toxoplasma*
657 *gondii* or (2) presence of that protein in either the isoprenoid biosynthesis or fatty acid
658 biosynthesis/utilization pathways. To generate a negative control list of potential false positives,
659 nuclear-encoded proteins (excluding pseudogenes) predicted to contain a signal peptide were
660 identified as above and 451 of these proteins were designated as negative controls based on GO
661 terms, annotations, and the published literature.

662

663 **Feature extraction for neural network**

664 To generate the positive training set for PlastNN, we took the combined list of previously known
665 apicoplast proteins (S2 Table) and apicoplast proteins identified by BioID (S1 Table) and
666 removed proteins that (1) were likely false positives based on our negative control list (S2 Table)
667 or published localization data; (2) were likely targeted to the apicoplast without the canonical
668 bipartite *N*-terminal leader sequence; or (3) did not contain a predicted signal peptide based on
669 the SignalP 3.0 *D*-score. This yielded a final positive training set of 205 proteins (S4 Table). The
670 negative training set was the previously generated list of known non-apicoplast proteins (S2

671 Table). The test set for PlastNN consisted of 450 proteins predicted to have a signal peptide by
672 the SignalP 3.0 *D*-score that were not in the positive or negative training sets.
673 For each protein in our training and test sets, we took the 50 amino acids immediately
674 after the end of the predicted signal peptide (according to the SignalP 3.0 *C*-score) and calculated
675 the frequency of each of the 20 amino acids in this sequence. The length of 50 amino acids was
676 chosen empirically by trying lengths from 20-100; highest accuracy was obtained using 50.
677 Scaled FPKM values at 8 time points during intraerythrocytic development were obtained from
678 published RNA-Seq [39]. By combining the amino acid frequencies with the 8 transcriptome
679 values, we represented each protein in our training and test sets by a feature vector of length 28.
680
681 **Neural network training and cross-validation**
682 To train the model, the 205 positive and 451 negative training examples were combined and
683 randomly shuffled. The training set was divided into 6 equal folds, each containing 109 or 110
684 examples. We trained models using 6-fold cross-validation; that is, we trained 6 separate models
685 with the same architecture, each using 5 of the 6 folds for training and then using the one
686 remaining fold as a cross-validation set to evaluate performance. Accuracy, sensitivity,
687 specificity, NPV, and PPV are calculated on this cross-validation set. The final reported values
688 of accuracy, sensitivity, specificity, NPV, and PPV are the average and standard deviation over
689 all 6 models. When predicting on the test set, the final predictions are generated by a majority
690 vote of all 6 models.
691 Neural networks were trained using the RMSProp optimization algorithm with a learning
692 rate of 0.0001. Tensorflow version 1.4.1 was used to build and train the neural network. Logistic

693 regression on the same dataset was carried out using the caret package (version 6.0-77) in R
694 version 3.3.3.
695
696 **Analyses of Apicoplast Proteome Datasets**
697 The BioID apicoplast proteome and the predicted proteomes from PATS, PlasmoAP, ApicoAP,
698 and PlastNN were analyzed according to the following formulae:
699 Accuracy = (TP + TN)/(TP + FP + TN + FN)
700 Sensitivity = TP/(TP + FN)
701 Specificity = TN/(TN + FP)
702 Negative Predictive Value (NPV) = TN/(TN + FN)
703 Positive Predictive Value (PPV) = TP/(TP + FP)
704 Abbreviations: TP, true positive; TN, true negative; FP, false positive; FN, false negative.
705 Because none of the 451 negative control proteins from the original list (S2 Table) were
706 identified in our 187-protein BioID proteome, we manually inspected the BioID list, identified 5
707 likely false positives, and added these to the negative control list for the purposes of analyses
708 presented in Fig 2 and S2 Fig.
709
710 **Protein Novelty Analysis**
711 Proteins in the apicoplast proteome were manually categorized for having a potentially novel
712 function based on PlasmoDB version 33 (released June 30, 2017) gene product annotations.
713 Gene products with annotations that could clearly assign a given protein to an established
714 cellular pathway were labeled as “Known Pathway;” gene products with a descriptive annotation
715 that did not clearly suggest a cellular pathway were labeled as “Annotated Gene Product,

716 Unknown Function;" and gene products that explicitly contained the words "unknown function"
717 were labeled as "Unknown Function."
718
719 **OrthoMCL Orthology Analysis**
720 To analyze the conservation of candidate apicoplast proteins identified by apicoplast BioID,
721 OrthoMCL ortholog group IDs were obtained from PlasmoDB. Based on OrthoMCL version 5
722 (released July 23, 2015), each ortholog group was then categorized as being present only in
723 *Plasmodium* spp., only in Apicomplexa, or present in at least one organism outside of the
724 Apicomplexa.
725
726 **Gene Essentiality Analysis**
727 Genome-scale essentiality data for *P. berghei* or *P. falciparum* genes were accessed from the
728 original manuscripts [41, 42].
729
730 **Selection of Candidates for Experimental Localization**
731 To facilitate molecular cloning, proteins identified by BioID or PlastNN were candidates for
732 GFP tagging only if their corresponding gene sizes were less than 2 kb. With the exception of
733 ROM7, which was selected based on the biological interest of rhomboid proteases, we focused
734 on localizing conserved *Plasmodium* genes of unknown function due to interest in functional
735 characterization of the *Plasmodium* genome. PF3D7_1472800, PF3D7_0521400, and
736 PF3D7_0721100 (all from the BioID list) were chosen due to their diverse apicoplast:ER
737 enrichment rankings in the BioID list (S1 Table; PF3D7_1472800 ranked number 52/187 and
738 was identified exclusively in BioID-Ap samples; PF3D7_0521400 ranked number 131/187 and

was found in both samples but was enriched nearly 400-fold in BioID-Ap samples; and PF3D7_0721100 ranked 184/187 and was enriched in BioID-Ap samples only slightly above our 5-fold cutoff). From the PlastNN list, PF3D7_1349900 and PF3D7_1330100 were selected solely based on being proteins of unknown function with small gene sizes. Because of the small sample sizes of proteins selected for GFP-tagging and the non-random nature of selecting candidates, we note that the results of our experimental validation should not be extrapolated to be representative of the PPVs of the BioID and PlastNN datasets as a whole.

Parasite Growth Time Courses

Sorbitol-synchronized ABCB7 and ABCF1 TetR-DOZI parasites were washed multiple times to remove residual ATc and were returned to culture medium containing 500 nM ATc, 200 µM IPP (Isoprenoids, LLC), or no supplements. Samples for growth assays, DNA isolation, or western blotting were harvested every other day when the majority of parasites were trophozoites and schizonts. For growth assays, parasites were fixed in 1% paraformaldehyde in PBS and were stored at 4°C until completion of the time course. Samples were then stained with 50 nM YOYO-1 and parasitemia was analyzed on a BD Accuri C6 flow cytometer. Samples for DNA isolation and western blotting were treated with 0.1% saponin in PBS to release parasites from the erythrocyte, washed in PBS, and stored at -80°C until analysis.

Quantitative PCR

Total parasite DNA was isolated from time course samples using the DNeasy Blood & Tissue Kit (Qiagen). Quantitative PCR was performed using Power SYBR Green PCR Master Mix (Thermo Fisher) with primers CHT1 F and CHT1 R targeting the nuclear gene chitinase or TufA

762 F and TufA R targeting the apicoplast gene elongation factor Tu (0.15 µM final concentration
763 each primer) [46]. Quantitative PCR was performed on an Applied Biosystems 7900HT Real-
764 Time PCR System with the following thermocycling conditions: Initial denaturation 95°C/10
765 minutes; 35 cycles of 95°C/1 minute, 56°C/1 minute; 65°C/1 minute; final extension 65°C/10
766 minutes. Relative quantification of each target was performed using the $\Delta\Delta C_t$ method.
767
768 **Statistics**
769 95% confidence intervals were determined using the GraphPad QuickCalc for confidence
770 interval of a proportion via the modified Wald method
771 (https://www.graphpad.com/quickcalcs/confInterval1/). Two-way ANOVAs were performed in
772 GraphPad Prism version 7.04.
773
774 **Acknowledgements**
775 We thank Jacquin Niles for providing the NF54$^{Cas9+T7\ Polymerase}$ cell line and pSN054-V5 plasmid,
776 Sean Prigge for α-*Pf*ACP antibody, Sebastian Mikolajczak and Stefan Kappe for α-*Pf*BiP
777 antibody, and Walid Houry for α-*Pf*ClpP antibody. We also thank Julian Lutze for assistance
778 with molecular cloning of candidate apicoplast genes.
779
780 **Author Contributions**
781 Conceptualization, M.J.B. and E.Y.; Software, A.L., S.J.W., A.J., S.Z., X.W., J.Z. and S.A.R.;
782 Investigation, M.J.B., and S.G.; Resources, L.Z., J.E.E., and S.A.R.; Writing – Original Draft,
783 M.J.B. and E.Y.; Writing – Review & Editing, M.J.B., S.G., L.Z., A.L., S.W.J., A.J., S.Z., X.W.,

784 S.A.R., J.Z., J.E.E., and E.Y.; Supervision, E.Y., J.E.E., and J.Z.; Funding Acquisition, E.Y.,

785 J.E.E., J.Z., and S.A.R.

786

787 **Figures and Legends**

788

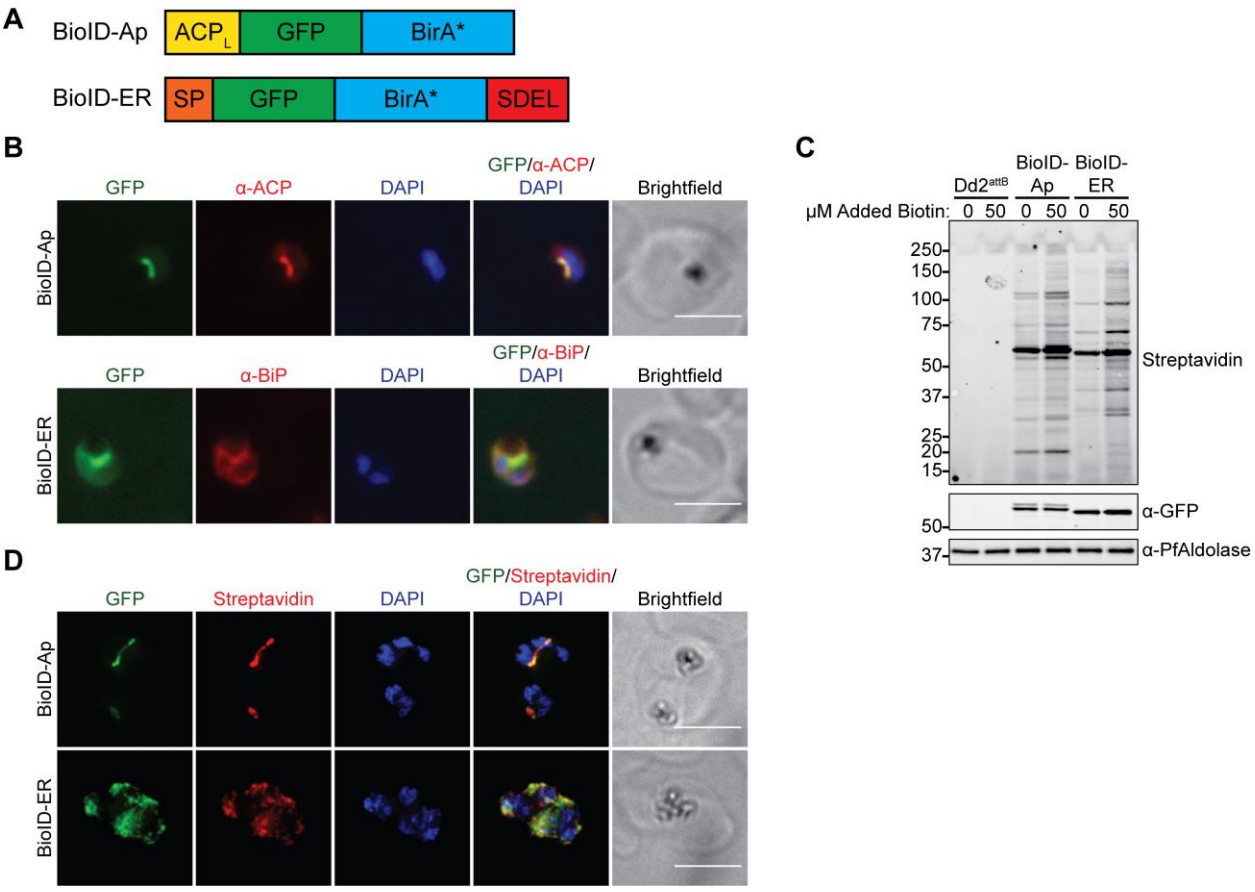


Fig 1. The promiscuous biotin ligase BirA* biotinylates proteins in the *P. falciparum* apicoplast and ER. (A) Schematic (not to scale) of constructs for apicoplast- and ER-targeting of GFP-BirA*. ACP_L, ACP leader sequence; SP, signal peptide; SDEL, ER-retention motif. (B) Fixed-cell imaging of BioID-Ap and BioID-ER parasites stained with antibodies raised against the apicoplast marker ACP or the ER marker BiP, respectively. Scale bars, 5 μm. (C) Western blot of untreated and biotin-labeled $Dd2^{attB}$, BioID-Ap, and BioID-ER parasites. (D) Fixed-cell imaging of biotinylated proteins in biotin-labeled BioID-Ap and BioID-ER parasites. Scale bars, 5 μm.

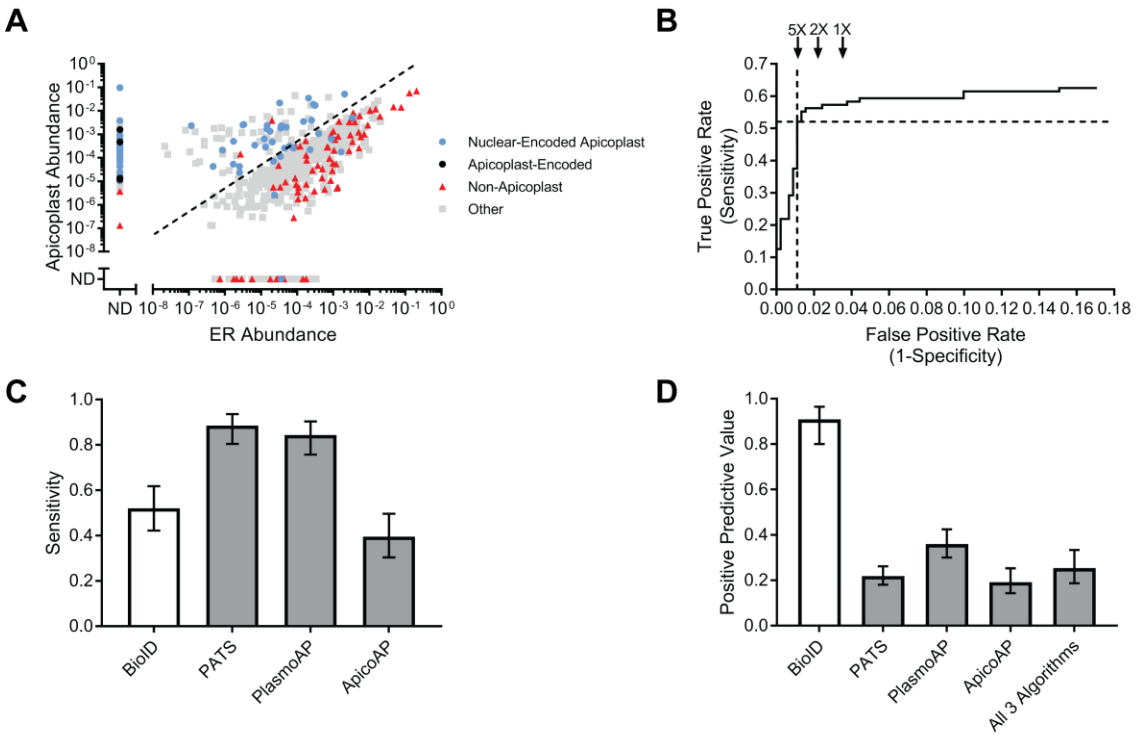


Fig 2. Accurate, unbiased identification of apicoplast proteins using BioID. (A) Abundances of 728 proteins identified by mass spectrometry in BioID-Ap and BioID-ER parasites. Protein abundances were calculated by summing the total MS1 area of all matched peptides for a given protein and normalized by the total summed intensity of all *P. falciparum* peptides matched. Dotted line represents 5-fold apicoplast:ER enrichment. ND, not detected. (B) ROC curve used to identify the apicoplast:ER enrichment that maximized true positives while minimizing false positives. Dotted lines denote the sensitivity and false positive rate of the 5-fold cutoff used. False positive rates for hypothetical 2-fold and 1-fold enrichments are shown for reference. (C) Sensitivities of BioID, PATS, PlasmoAP, and ApicoAP based on identification of 96 known apicoplast proteins. (D) PPV of BioID, PATS, PlasmoAP, ApicoAP, and a dataset consisting of proteins predicted to localize to the apicoplast by all three bioinformatic algorithms. Calculated as the number of true positives divided by the total number of true positives and false positives. Error bars in (C) and (D) represent 95% confidence intervals.

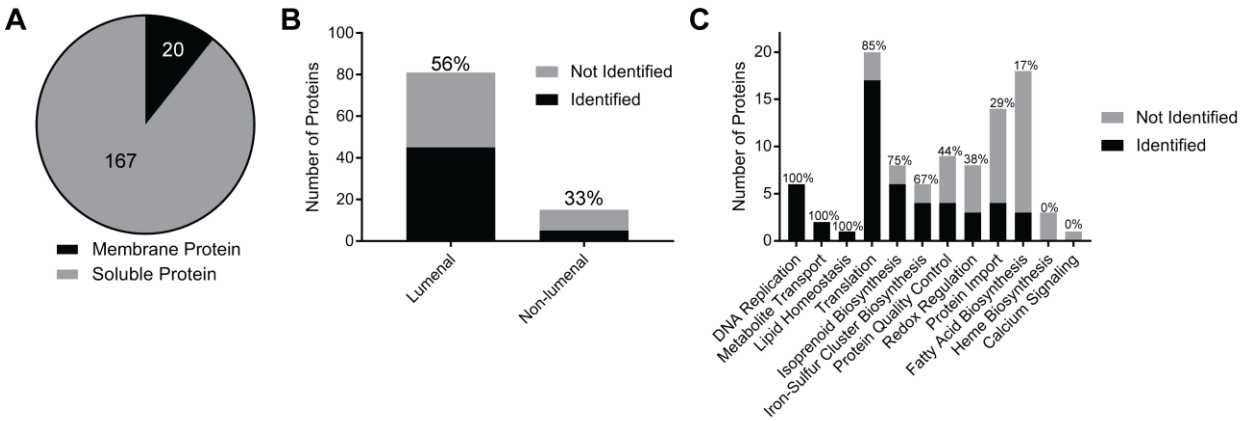
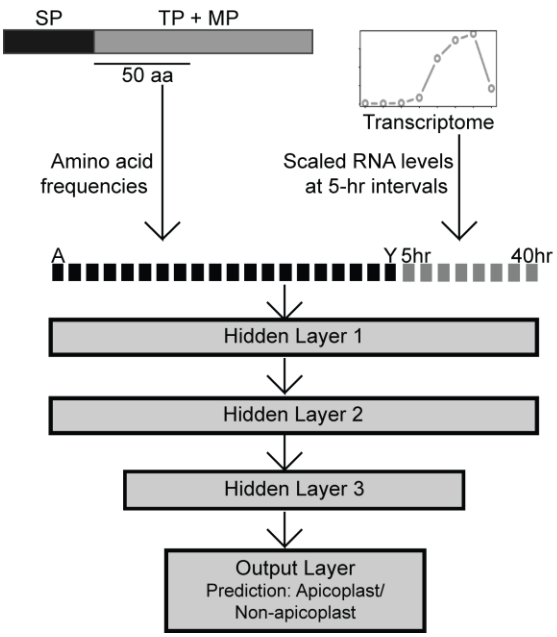


Fig 3. Diversity of protein labeling by apicoplast BioID. (A) Fraction of proteins identified by apicoplast BioID that are predicted to localize to a membrane. Proteins were considered "membrane" if they had at least one transmembrane domain annotated in PlasmoDB ending >80 amino acids from the annotated *N*-terminus. (B) Number of lumenal and non-lumenal positive controls identified. Percentages above bars indicate the percentage of known proteins from each category identified. (C) Number of proteins from established apicoplast pathways identified. Percentages above bars indicate the percentage of known proteins from each pathway identified.

A



B

Model	1	2	3	4	5	6	Mean	SD
Accuracy	97%	94%	94%	95%	99%	99%	**96%**	**2%**
Sensitivity	100%	91%	89%	90%	100%	97%	**95%**	**5%**
Specificity	96%	95%	97%	99%	99%	100%	**98%**	**2%**
NPV	100%	96%	95%	94%	100%	99%	**97%**	**3%**
PPV	91%	89%	94%	97%	97%	100%	**95%**	**4%**

C

Algorithm	Accuracy	Sensitivity	Specificity	NPV	PPV
PATS	53%	97%	33%	96%	40%
PlasmoAP	77%	95%	69%	97%	58%
ApicoAP	66%	70%	65%	83%	48%
PlastNN	**96%**	**95%**	**98%**	**97%**	**95%**

Fig 4. Improved prediction of apicoplast proteins using the PlastNN algorithm. (A) Schematic of the PlastNN algorithm. For each signal peptide-containing protein, a region of 50 amino acids immediately following the signal peptide cleavage site was selected and the frequencies of the 20 canonical amino acids in this region were calculated, resulting in a vector of length 20. Scaled RNA levels of the gene encoding the protein at 8 time points were added, resulting in a 28-dimensional vector representing each protein. This was used as input to train a neural network with 3 hidden layers, resulting in a prediction of whether the protein is targeted to the apicoplast or not. (B) Table showing the performance of the 6 models in PlastNN. Each model was trained on 5/6th of the training set and cross-validated on the remaining 1/6th. Values shown are accuracy, sensitivity, specificity, NPV, and PPV on the cross-validation set. The final values reported are the average and standard deviation over all 6 models. (C) Comparison of accuracy, sensitivity, specificity, NPV, and PPV for three previous algorithms and PlastNN.

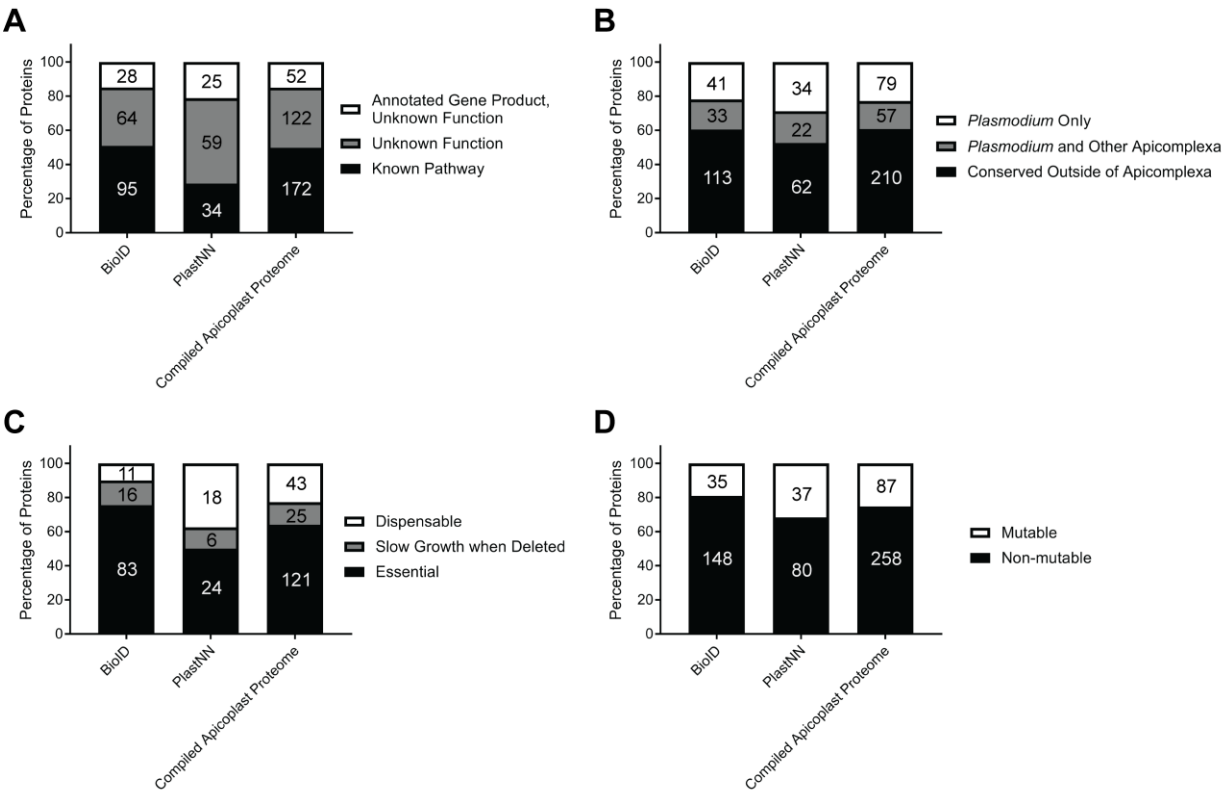
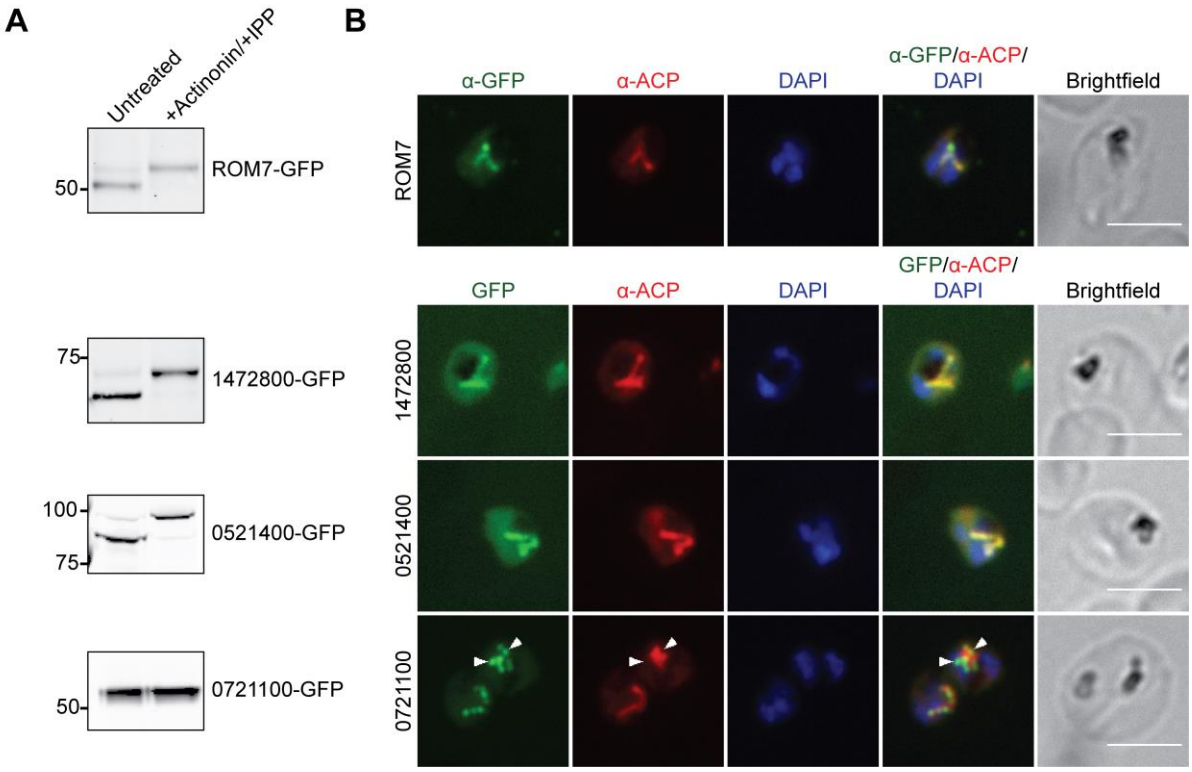


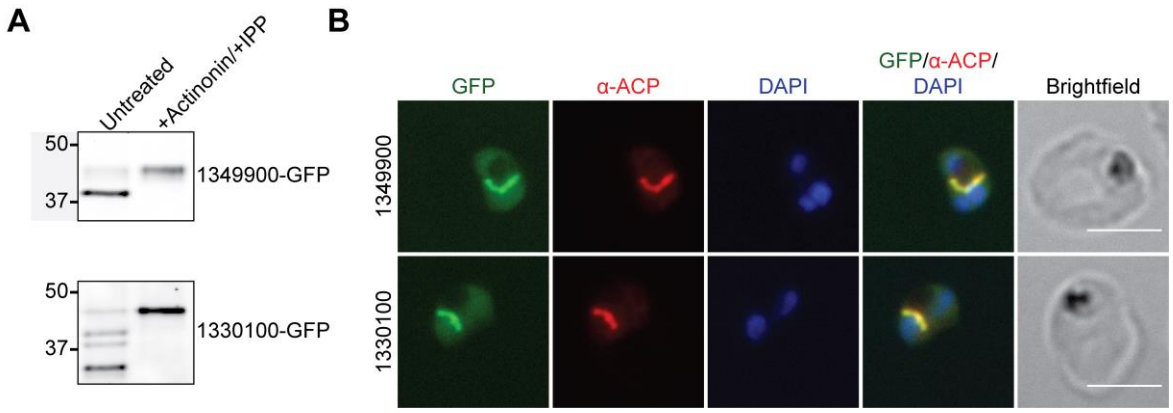
Fig 5. Apicoplast BioID identifies novel and essential proteins. (A) Percentage of proteins identified that have 1) annotated gene products but unknown function, 2) gene products annotated explicitly with "unknown function," or 3) annotated gene products and function in a known cellular pathway. (B) Percentage of proteins identified that are *Plasmodium*- or Apicomplexa-specific based on OrthoMCL orthology analysis. (C) Percentage of proteins identified that are essential, cause slow growth when deleted, or are dispensable based on PlasmoGEM essentiality data of *P. berghei* orthologs [41]. (D) Percentage of proteins identified that were classified as mutable or non-mutable based on genome-scale transposon mutagenesis in *P. falciparum* [42]. In each panel, absolute numbers of proteins are indicated within bars.



847

848 **Fig 6. Localization of candidate apicoplast proteins identified by BioID.** (A) Transit peptide
849 processing assay for *C*-terminally GFP-tagged candidates. Ring-stage parasites were either
850 untreated or treated with 10 µM actinonin/200 µM IPP for 3 days and protein processing was
851 assessed by western blot. (B) Fixed-cell imaging of GFP-tagged candidates in parasites stained
852 with an antibody raised against the apicoplast marker ACP. ROM7-GFP-expressing parasites
853 were also stained with anti-GFP antibody due to low signal from intrinsic GFP fluorescence in
854 fixed cells. Arrowheads indicate regions where PF3D7_0721100-GFP puncta appear adjacent to
855 as opposed to co-localizing with ACP. Scale bars, 5 µm.

856



857

858 **Fig 7. Localization of candidate apicoplast proteins identified by PlastNN.** (A) Transit

859 peptide processing assay for *C*-terminally GFP-tagged candidates. Ring-stage parasites were

860 either untreated or treated with 10 µM actinonin/200 µM IPP for 3 days and protein processing

861 was assessed by western blot. (B) Fixed-cell imaging of GFP-tagged candidates in parasites

862 stained with an antibody raised against the apicoplast marker ACP. Scale bars, 5 µm.

863

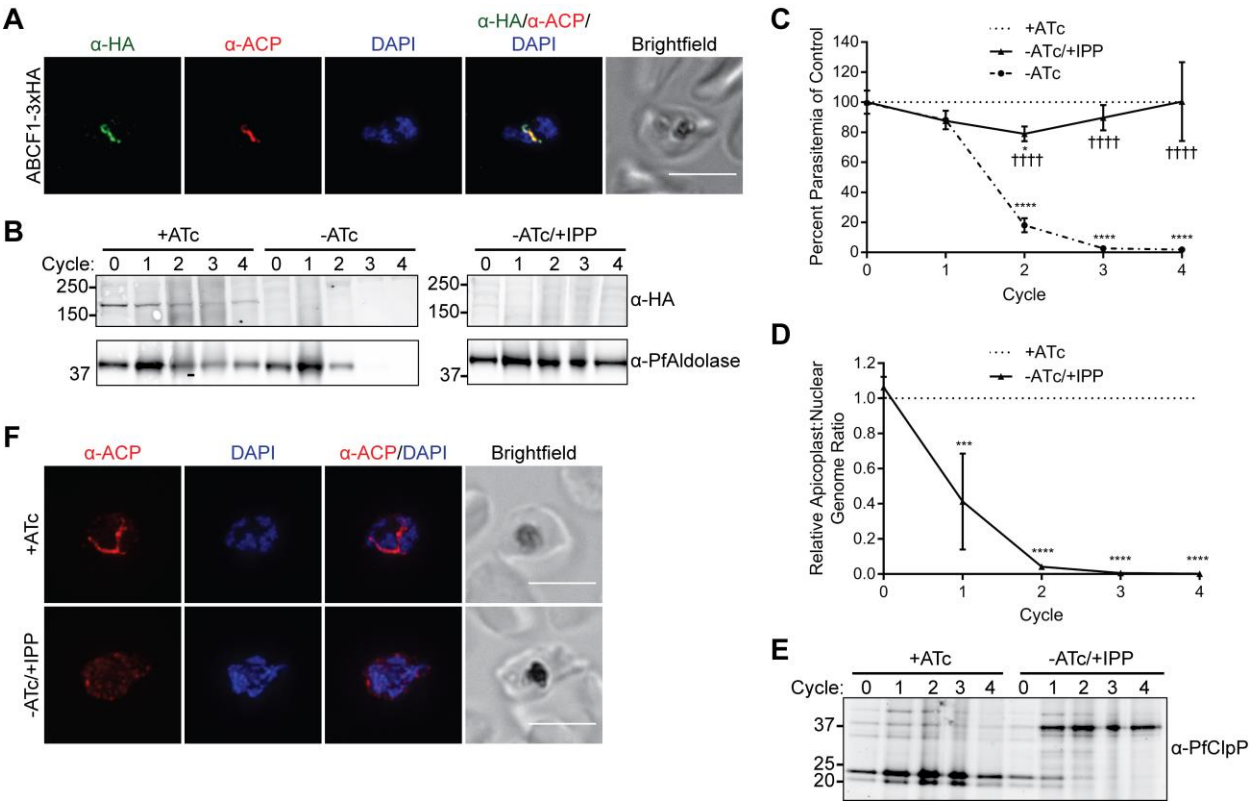


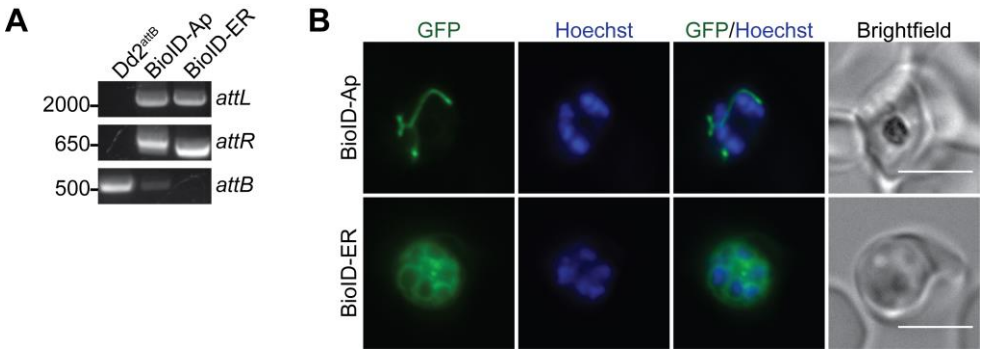
Fig 8. ABCF1 is an essential apicoplast protein required for organelle biogenesis. (A) Fixed-cell imaging of ABCF1-3xHA knockdown parasites stained with antibodies raised against the HA tag and the apicoplast marker ACP. Scale bar, 5 μm. (B-F) ABCF1-3xHA knockdown parasites were grown in the presence of ATc (+ATc), the absence of ATc (-ATc), or the absence of ATc with IPP supplementation (-ATc/+IPP) for 4 growth cycles. (B) Western blot of ABCF1-3xHA expression. (C) Parasite growth. At each time point, data are normalized to the untreated (+ATc) control. Error bars represent standard deviation from the mean of two biological replicates. *$P < 0.05$, ****$P < 0.0001$ compared to untreated control, ††††$P < 0.0001$ compared to -ATc condition, repeated measures two-way ANOVA with Tukey's multiple comparisons test. (D) Apicoplast:nuclear genome ratio. At each time point, data are normalized to the untreated (+ATc) control. Error bars represent standard deviation from the mean of two biological replicates, each performed in technical triplicate. ***$P < 0.001$, ****$P < 0.0001$, repeated

877 measures two-way ANOVA with Sidak's multiple comparisons test. (E) Western blot of ClpP

878 processing. (F) Fixed-cell imaging showing ACP localization after 2 cycles of knockdown. Scale

879 bars, 5 µm.

880



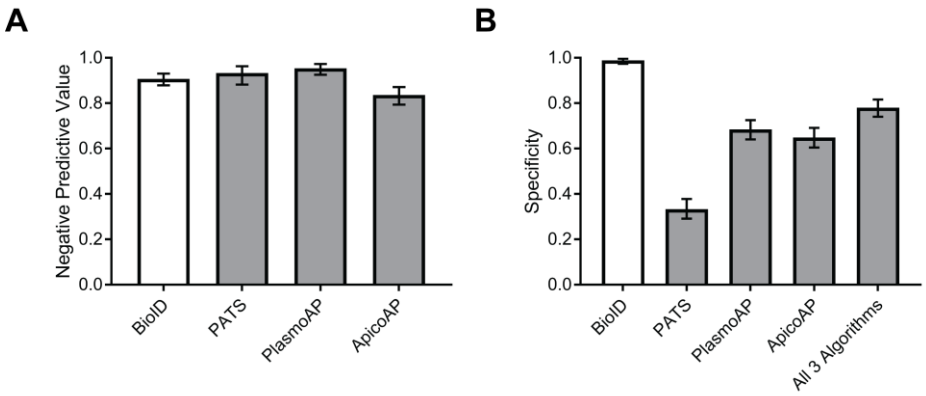
881

882 **S1 Fig. Integration and expression of BioID-Ap and BioID-ER constructs in Dd2^{attB}**

883 **parasites.** (A) PCR products showing integrated *attL* and *attR* sites or unintegrated *attB* site. (B)

884 Live-cell imaging of Hoechst-stained BioID-Ap and BioID-ER parasites. Scale bars, 5 µm.

885



886

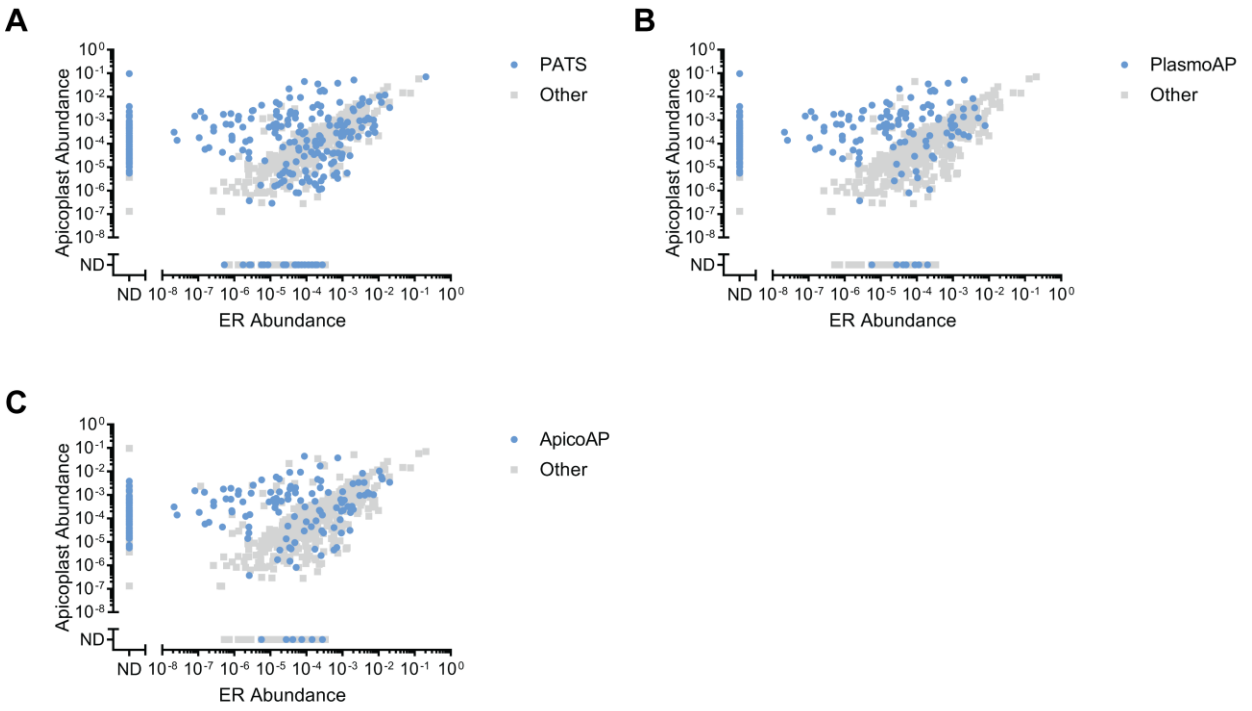
887 **S2 Fig. Comparison of negative predictive values and specificities of apicoplast BioID and**

888 **bioinformatic prediction algorithms.** (A) Negative predictive value (NPV) and (B) specificity

889 of apicoplast BioID, PATS, PlasmoAP, and ApicoAP. Error bars represent 95% confidence

890 intervals.

891



892

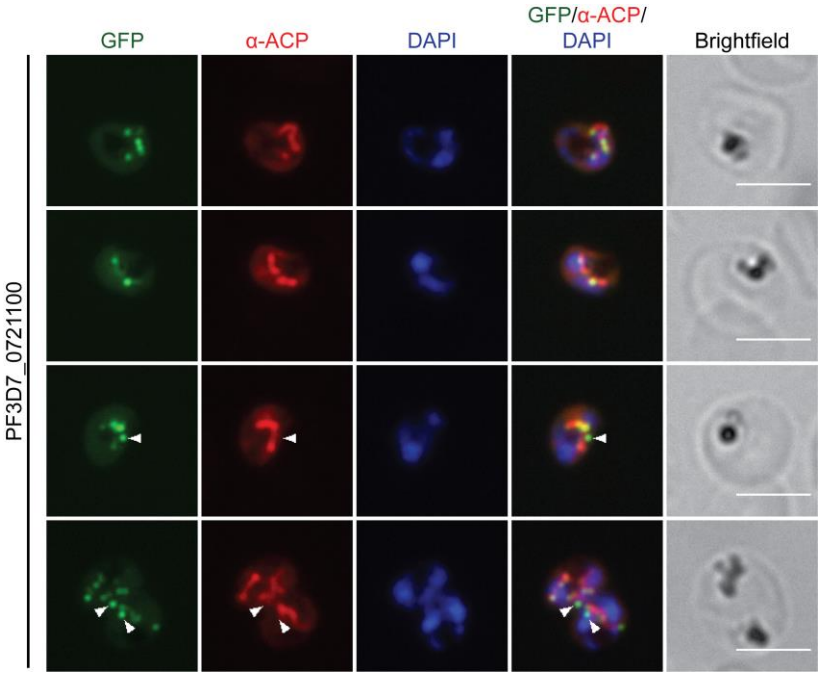
893 **S3 Fig. Bioinformatically predicted apicoplast proteins are not clearly distinguishable**

894 **based on apicoplast:ER abundance ratio.** Proteins predicted to localize to the apicoplast by

895 (A) PATS, (B) PlasmoAP, or (C) ApicoAP are highlighted in each graph. Data points are

896 identical to those in Fig 2A.

897



898

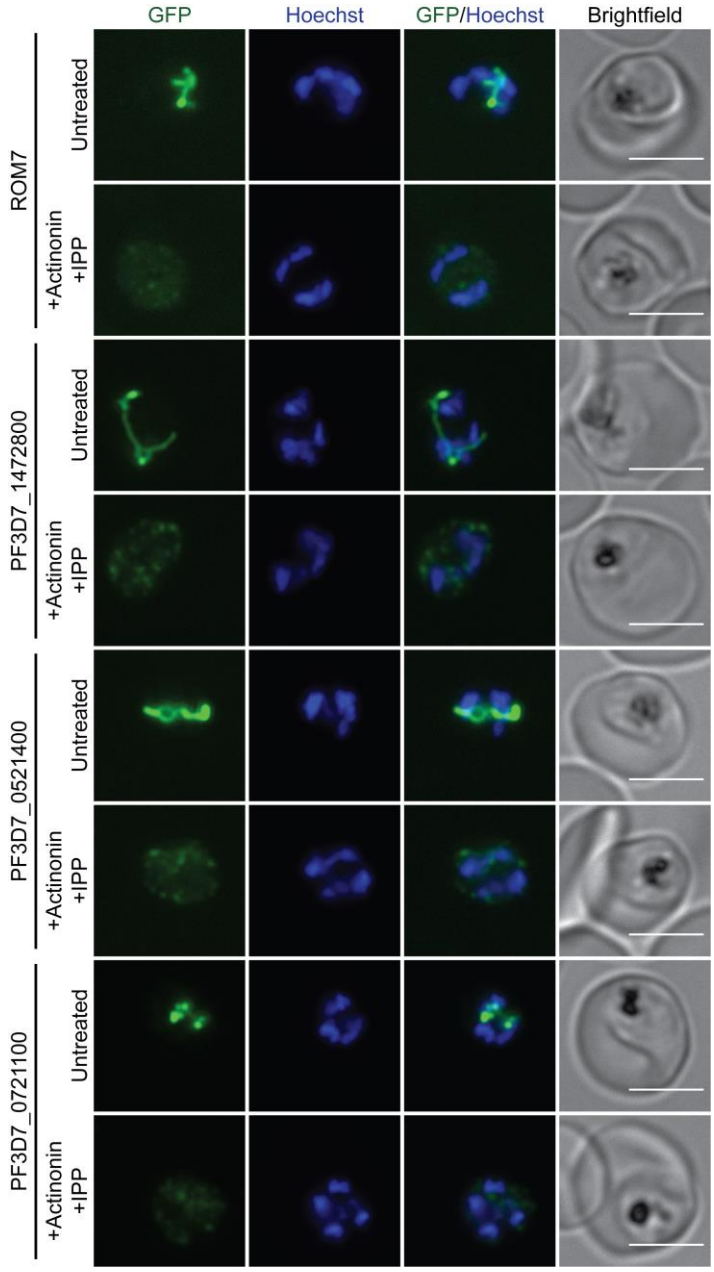
899 **S4 Fig. Additional fixed-cell images of PF3D7_0721100-GFP localization.** PF3D7_0721100-

900 GFP parasites were stained with an antibody against the apicoplast marker ACP. Arrowheads

901 indicate regions where PF3D7_0721100-GFP puncta appear adjacent to as opposed to co-

902 localizing with ACP. Scale bars, 5 μm.

903



904

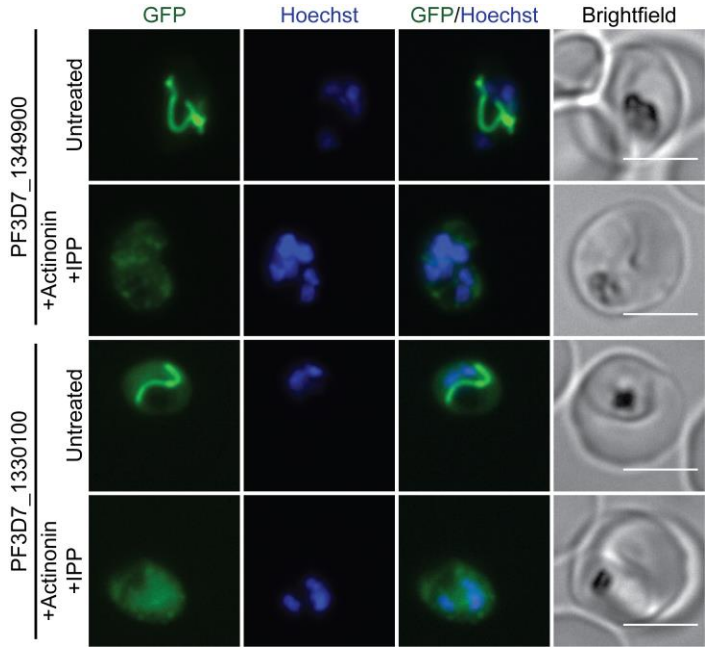
905 **S5 Fig. Live-cell imaging of candidate apicoplast proteins identified by BioID.** Parasites

906 expressing *C*-terminally GFP-tagged candidate proteins from apicoplast BioID were either

907 untreated (apicoplast-intact) or treated with 10 µM actinonin/200 µM IPP (apicoplast-disrupted)

908 for 3 days prior to imaging. Scale bars, 5 µm.

909



910

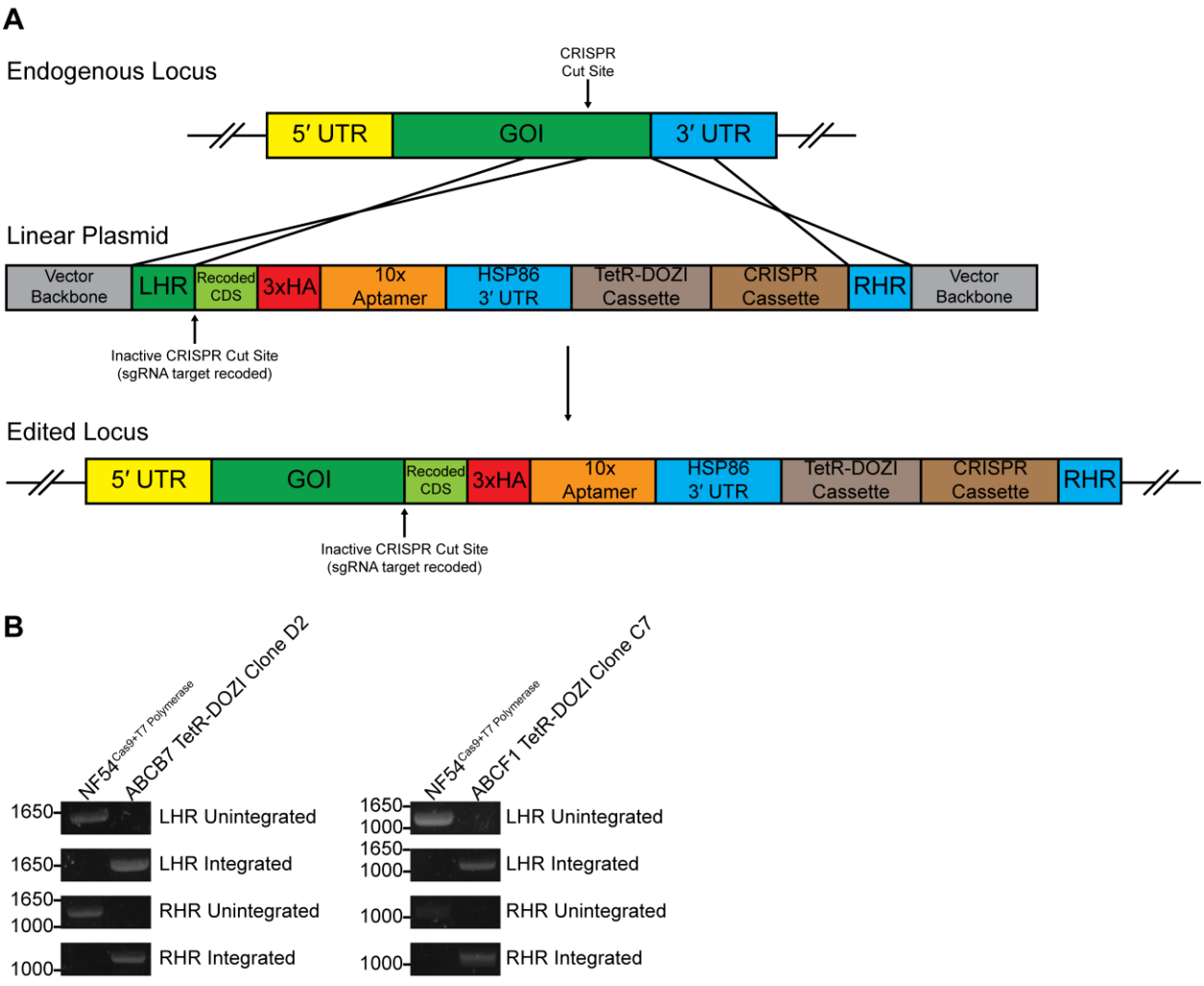
911 **S6 Fig. Live-cell imaging of candidate apicoplast proteins identified by PlastNN.** Parasites

912 expressing *C*-terminally GFP-tagged candidate proteins from PlastNN were either untreated

913 (apicoplast-intact) or treated with 10 µM actinonin/200 µM IPP (apicoplast-disrupted) for 3 days

914 prior to imaging. Scale bars, 5 µm.

915



916

917 **S7 Fig. Generation of ABCB7 and ABCF1 TetR-DOZI conditional knockdown cell lines.**

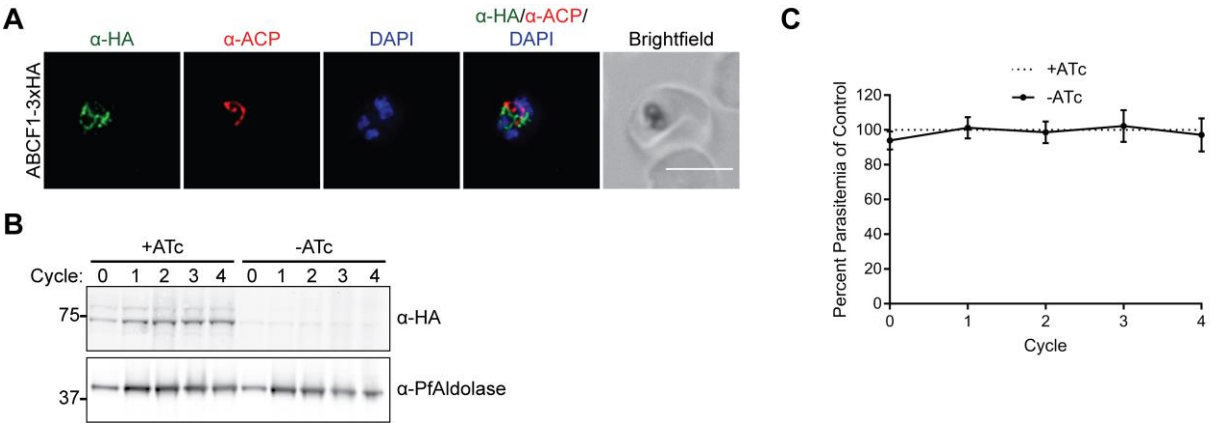
918 (A) Schematic of CRISPR-Cas9-based endogenous editing to generate conditional knockdown

919 cell lines. GOI, gene of interest; LHR, left homology region; RHR, right homology region. (B)

920 PCR products showing integrated or unintegrated LHR and RHR sites in parental NF54$^{Cas9+T7}$

921 ^{Polymerase} or clonal genome-edited parasites.

922



S8 Fig. ABCB7 is a probable non-apicoplast protein for which knockdown does not cause growth inhibition. (A) Fixed-cell imaging of ABCB7-3xHA knockdown parasites stained with antibodies raised against the HA tag and the apicoplast marker ACP. Scale bar, 5 μm. (B-C) ABCB7-3xHA knockdown parasites were grown in the presence of ATc (+ATc) or the absence of ATc (-ATc) for 4 growth cycles. (B) Western blot of ABCB7-3xHA expression. (C) Parasite growth. At each time point, data are normalized to the untreated (+ATc) control. Error bars represent standard deviation from the mean of two biological replicates.

S1 Table. Abundances of 728 *P. falciparum* proteins identified by mass spectrometry in ≥2 biological replicates and with ≥2 unique peptides in at least one mass spectrometry run.

S2 Table. Positive and negative control apicoplast proteins used in this study.

S3 Table. Proteins predicted to localize to the apicoplast by PATS, PlasmoAP, and ApicoAP.

S4 Table. Positive training set used to develop PlastNN.

S5 Table. Layer dimensions for PlastNN neural network.

S6 Table. Performance of different models in cross-validation.

S7 Table. Results of PlastNN prediction algorithm.

S8 Table. Compiled list of 346 candidate apicoplast proteins based on localization in the published literature, BioID, and PlastNN.

S9 Table. Summary of BioID and PlastNN candidate localization data from this study and Sayers et al.

S10 Table. Primer and gBlock sequences used in this study.

S1 Data. Spreadsheet containing numerical data for Figs 2A, 2B, 2C, 2D, 3A, 3B, 3C, 5A, 5B, 5C, 5D, 8C, 8D, S2A, S2B, S7C.

References

1. Aurrecoechea C, Barreto A, Basenko EY, Brestelli J, Brunk BP, Cade S, et al. EuPathDB: the eukaryotic pathogen genomics database resource. Nucleic Acids Res. 2017;45(D1):D581-91. doi: 10.1093/nar/gkw1105. PMID: 27903906; PMCID: PMC5210576.

2. McFadden GI, Reith ME, Munholland J, Lang-Unnasch N. Plastid in human parasites. Nature. 1996;381(6582):482. doi: 10.1038/381482a0. PMID: 8632819.

3. Kohler S, Delwiche CF, Denny PW, Tilney LG, Webster P, Wilson RJ, et al. A plastid of probable green algal origin in Apicomplexan parasites. Science. 1997;275(5305):1485-9. PMID: 9045615.

4. van Dooren GG, Striepen B. The algal past and parasite present of the apicoplast. Annu Rev Microbiol. 2013;67:271-89. doi: 10.1146/annurev-micro-092412-155741. PMID: 23808340.

5. Spork S, Hiss JA, Mandel K, Sommer M, Kooij TW, Chu T, et al. An unusual ERAD-like complex is targeted to the apicoplast of *Plasmodium falciparum*. Eukaryot Cell. 2009;8(8):1134-45. doi: 10.1128/ec.00083-09. PMID: 19502583; PMCID: PMC2725561.

6. Kalanon M, Tonkin CJ, McFadden GI. Characterization of two putative protein translocation components in the apicoplast of *Plasmodium falciparum*. Eukaryot Cell. 2009;8(8):1146-54. doi: 10.1128/ec.00061-09. PMID: 19502580; PMCID: PMC2725556.

7. Agrawal S, van Dooren GG, Beatty WL, Striepen B. Genetic evidence that an endosymbiont-derived endoplasmic reticulum-associated protein degradation (ERAD) system functions in import of apicoplast proteins. J Biol Chem. 2009;284(48):33683-91. doi: 10.1074/jbc.M109.044024. PMID: 19808683; PMCID: PMC2785210.

8. Agrawal S, Chung DW, Ponts N, van Dooren GG, Prudhomme J, Brooks CF, et al. An apicoplast localized ubiquitylation system is required for the import of nuclear-encoded plastid proteins. PLOS Pathog. 2013;9(6):e1003426. doi: 10.1371/journal.ppat.1003426. PMID: 23785288; PMCID: PMC3681736.

9. Ralph SA, van Dooren GG, Waller RF, Crawford MJ, Fraunholz MJ, Foth BJ, et al. Tropical infectious diseases: metabolic maps and functions of the *Plasmodium falciparum* apicoplast. Nat Rev Microbiol. 2004;2(3):203-16. doi: 10.1038/nrmicro843. PMID: 15083156.

10. Sheiner L, Vaidya AB, McFadden GI. The metabolic roles of the endosymbiotic organelles of *Toxoplasma* and *Plasmodium* spp. Curr Opin Microbiol. 2013;16(4):452-8. doi: 10.1016/j.mib.2013.07.003. PMID: 23927894; PMCID: PMC3767399.

11. Jomaa H, Wiesner J, Sanderbrand S, Altincicek B, Weidemeyer C, Hintz M, et al. Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. Science. 1999;285(5433):1573-6. PMID: 10477522.

12. Dahl EL, Shock JL, Shenai BR, Gut J, DeRisi JL, Rosenthal PJ. Tetracyclines specifically target the apicoplast of the malaria parasite *Plasmodium falciparum*. Antimicrob Agents Chemother. 2006;50(9):3124-31. doi: 10.1128/aac.00394-06. PMID: 16940111; PMCID: PMC1563505.

13. Dahl EL, Rosenthal PJ. Multiple antibiotics exert delayed effects against the *Plasmodium falciparum* apicoplast. Antimicrob Agents Chemother. 2007;51(10):3485-90. doi: 10.1128/aac.00527-07. PMID: 17698630; PMCID: PMC2043295.

14. Goodman CD, Su V, McFadden GI. The effects of anti-bacterials on the malaria parasite *Plasmodium falciparum*. Mol Biochem Parasitol. 2007;152(2):181-91. doi: 10.1016/j.molbiopara.2007.01.005. PMID: 17289168.

15. Stanway RR, Witt T, Zobiak B, Aepfelbacher M, Heussler VT. GFP-targeting allows visualization of the apicoplast throughout the life cycle of live malaria parasites. Biol Cell. 2009;101(7):415-30. doi: 10.1042/bc20080202. PMID: 19143588.

16. Amberg-Johnson K, Hari SB, Ganesan SM, Lorenzi HA, Sauer RT, Niles JC, et al. Small molecule inhibition of apicomplexan FtsH1 disrupts plastid biogenesis in human pathogens. eLife. 2017;6:e29865. doi: 10.7554/eLife.29865. PMID: 28826494; PMCID: PMC5576918.

17. 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. PMID: 11738814.

18. Foth BJ, Ralph SA, Tonkin CJ, Struck NS, Fraunholz M, Roos DS, et al. Dissecting apicoplast targeting in the malaria parasite *Plasmodium falciparum*. Science. 2003;299(5607):705-8. doi: 10.1126/science.1078599. PMID: 12560551.

19. Cilingir G, Broschat SL, Lau AO. ApicoAP: the first computational model for identifying apicoplast-targeted proteins in multiple species of Apicomplexa. PLOS ONE. 2012;7(5):e36598. doi: 10.1371/journal.pone.0036598. PMID: 22574192; PMCID: PMC3344922.

20. Roux KJ, Kim DI, Raida M, Burke B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. J Cell Biol. 2012;196(6):801-10. doi: 10.1083/jcb.201112098. PMID: 22412018; PMCID: PMC3308701.

21. Rhee HW, Zou P, Udeshi ND, Martell JD, Mootha VK, Carr SA, et al. Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. Science. 2013;339(6125):1328-31. doi: 10.1126/science.1230593. PMID: 23371551; PMCID: PMC3916822.

22. Morriswood B, Havlicek K, Demmel L, Yavuz S, Sealey-Cardona M, Vidilaseris K, et al. Novel bilobe components in *Trypanosoma brucei* identified using proximity-dependent biotinylation. Eukaryot Cell. 2013;12(2):356-67. doi: 10.1128/ec.00326-12. PMID: 23264645; PMCID: PMC3571296.

23. Chen AL, Kim EW, Toh JY, Vashisht AA, Rashoff AQ, Van C, et al. Novel components of the *Toxoplasma* inner membrane complex revealed by BioID. mBio. 2015;6(1):e02357-14. doi: 10.1128/mBio.02357-14. PMID: 25691595; PMCID: PMC4337574.

24. Nadipuram SM, Kim EW, Vashisht AA, Lin AH, Bell HN, Coppens I, et al. *In vivo* biotinylation of the *Toxoplasma* parasitophorous vacuole reveals novel dense granule proteins important for parasite growth and pathogenesis. mBio. 2016;7(4):e00808-16. doi: 10.1128/mBio.00808-16. PMID: 27486190; PMCID: PMC4981711.

25. Dang HQ, Zhou Q, Rowlett VW, Hu H, Lee KJ, Margolin W, et al. Proximity interactions among basal body components in *Trypanosoma brucei* identify novel regulators of basal body

biogenesis and inheritance. mBio. 2017;8(1):e02120-16. doi: 10.1128/mBio.02120-16. PMID: 28049148; PMCID: PMC5210500.

26. Chen AL, Moon AS, Bell HN, Huang AS, Vashisht AA, Toh JY, et al. Novel insights into the composition and function of the *Toxoplasma* IMC sutures. Cell Microbiol. 2017;19(4):e12678. doi: 10.1111/cmi.12678. PMID: 27696623; PMCID: PMC5909696.

27. Kehrer J, Frischknecht F, Mair GR. Proteomic analysis of the *Plasmodium berghei* gametocyte egressome and vesicular bioID of osmiophilic body proteins identifies merozoite TRAP-like protein (MTRAP) as an essential factor for parasite transmission. Mol Cell Proteomics. 2016;15(9):2852-62. doi: 10.1074/mcp.M116.058263. PMID: 27371728; PMCID: PMC5013303.

28. Khosh-Naucke M, Becker J, Mesen-Ramirez P, Kiani P, Birnbaum J, Frohlke U, et al. Identification of novel parasitophorous vacuole proteins in *P. falciparum* parasites using BioID. Int J Med Microbiol. 2018;308(1):13-24. doi: 10.1016/j.ijmm.2017.07.007. PMID: 28784333.

29. Schnider CB, Bausch-Fluck D, Bruhlmann F, Heussler VT, Burda PC. BioID reveals novel proteins of the *Plasmodium* parasitophorous vacuole membrane. mSphere. 2018;3(1):e00522-17. doi: 10.1128/mSphere.00522-17. PMID: 29404413; PMCID: PMC5784244.

30. Waller RF, Reed MB, Cowman AF, McFadden GI. Protein trafficking to the plastid of *Plasmodium falciparum* is via the secretory pathway. EMBO J. 2000;19(8):1794-802. doi: 10.1093/emboj/19.8.1794. PMID: 10775264; PMCID: PMC302007.

31. Nkrumah LJ, Muhle RA, Moura PA, Ghosh P, Hatfull GF, Jacobs WR, Jr., et al. Efficient site-specific integration in *Plasmodium falciparum* chromosomes mediated by mycobacteriophage Bxb1 integrase. Nat Methods. 2006;3(8):615-21. doi: 10.1038/nmeth904. PMID: 16862136; PMCID: PMC2943413.

32. Yu M, Kumar TR, Nkrumah LJ, Coppi A, Retzlaff S, Li CD, et al. The fatty acid biosynthesis enzyme FabI plays a key role in the development of liver-stage malarial parasites. Cell Host Microbe. 2008;4(6):567-78. doi: 10.1016/j.chom.2008.11.001. PMID: 19064257; PMCID: PMC2646117.

33. Vaughan AM, O'Neill MT, Tarun AS, Camargo N, Phuong TM, Aly AS, et al. Type II fatty acid synthesis is essential only for malaria parasite late liver stage development. Cell Microbiol. 2009;11(3):506-20. doi: 10.1111/j.1462-5822.2008.01270.x. PMID: 19068099; PMCID: PMC2688669.

34. Pei Y, Tarun AS, Vaughan AM, Herman RW, Soliman JM, Erickson-Wayman A, et al. *Plasmodium* pyruvate dehydrogenase activity is only essential for the parasite's progression from liver infection to blood infection. Mol Microbiol. 2010;75(4):957-71. doi: 10.1111/j.1365-2958.2009.07034.x. PMID: 20487290.

35. Nagaraj VA, Sundaram B, Varadarajan NM, Subramani PA, Kalappa DM, Ghosh SK, et al. Malaria parasite-synthesized heme is essential in the mosquito and liver stages and complements host heme in the blood stages of infection. PLOS Pathog. 2013;9(8):e1003522. doi: 10.1371/journal.ppat.1003522. PMID: 23935500; PMCID: PMC3731253.

36. Ke H, Sigala PA, Miura K, Morrisey JM, Mather MW, Crowley JR, et al. The heme biosynthesis pathway is essential for *Plasmodium falciparum* development in mosquito stage but not in blood stages. J Biol Chem. 2014;289(50):34827-37. doi: 10.1074/jbc.M114.615831. PMID: 25352601; PMCID: PMC4263882.

37. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLOS Biol. 2003;1(1):E5. doi: 10.1371/journal.pbio.0000005. PMID: 12929205; PMCID: PMC176545.

38. Sheiner L, Demerly JL, Poulsen N, Beatty WL, Lucas O, Behnke MS, et al. A systematic screen to discover and analyze apicoplast proteins identifies a conserved and essential protein import factor. PLOS Pathog. 2011;7(12):e1002392. doi: 10.1371/journal.ppat.1002392. PMID: 22144892; PMCID: PMC3228799.

39. Bartfai R, Hoeijmakers WA, Salcedo-Amaya AM, Smits AH, Janssen-Megens E, Kaan A, et al. H2A.Z demarcates intergenic regions of the *Plasmodium falciparum* epigenome that are dynamically marked by H3K9ac and H3K4me3. PLOS Pathog. 2010;6(12):e1001223. doi: 10.1371/journal.ppat.1001223. PMID: 21187892; PMCID: PMC3002978.

40. Chen F, Mackey AJ, Stoeckert CJ, Jr., Roos DS. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. 2006;34(Database issue):D363-8. doi: 10.1093/nar/gkj123. PMID: 16381887; PMCID: PMC1347485.

41. Bushell E, Gomes AR, Sanderson T, Anar B, Girling G, Herd C, et al. Functional profiling of a *Plasmodium* genome reveals an abundance of essential genes. Cell. 2017;170(2):260-72.e8. doi: 10.1016/j.cell.2017.06.030. PMID: 28708996; PMCID: PMC5509546.

42. Zhang M, Wang C, Otto TD, Oberstaller J, Liao X, Adapa SR, et al. Uncovering the essential genes of the human malaria parasite *Plasmodium falciparum* by saturation mutagenesis. Science. 2018;360(6388):eaap7847. doi: 10.1126/science.aap7847. PMID: 29724925.

43. Sayers CP, Mollard V, Buchanan HD, McFadden GI, Goodman CD. A genetic screen in rodent malaria parasites identifies five new apicoplast putative membrane transporters, one of which is essential in human malaria parasites. Cell Microbiol. 2018;20(1):e12789. doi: 10.1111/cmi.12789. PMID: 28902970.

44. Waller RF, Keeling PJ, Donald RG, Striepen B, Handman E, Lang-Unnasch N, et al. Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. Proc Natl Acad Sci U S A. 1998;95(21):12352-7. PMID: 9770490; PMCID: PMC22835.

45. van Dooren GG, Su V, D'Ombrain MC, McFadden GI. Processing of an apicoplast leader sequence in *Plasmodium falciparum* and the identification of a putative leader cleavage enzyme. J Biol Chem. 2002;277(26):23612-9. doi: 10.1074/jbc.M201748200. PMID: 11976331.

46. Yeh E, DeRisi JL. Chemical rescue of malaria parasites lacking an apicoplast defines organelle function in blood-stage *Plasmodium falciparum*. PLOS Biol. 2011;9(8):e1001138. doi: 10.1371/journal.pbio.1001138. PMID: 21912516; PMCID: PMC3166167.

47. Kerr ID. Sequence analysis of twin ATP binding cassette proteins involved in translational control, antibiotic resistance, and ribonuclease L inhibition. Biochem Biophys Res Commun. 2004;315(1):166-73. doi: 10.1016/j.bbrc.2004.01.044. PMID: 15013441.

48. Dean M, Annilo T. Evolution of the ATP-binding cassette (ABC) transporter superfamily in vertebrates. Annu Rev Genomics Hum Genetics. 2005;6:123-42. doi: 10.1146/annurev.genom.6.080604.162122. PMID: 16124856.

49. Rijpma SR, van der Velden M, Annoura T, Matz JM, Kenthirapalan S, Kooij TW, et al. Vital and dispensable roles of *Plasmodium* multidrug resistance transporters during blood- and mosquito-stage development. Mol Microbiol. 2016;101(1):78-91. doi: 10.1111/mmi.13373. PMID: 26991313.

50. Goldfless SJ, Wagner JC, Niles JC. Versatile control of *Plasmodium falciparum* gene expression with an inducible protein-RNA interaction. Nat Commun. 2014;5:5329. doi: 10.1038/ncomms6329. PMID: 25370483; PMCID: PMC4223869.

51. Ganesan SM, Falla A, Goldfless SJ, Nasamu AS, Niles JC. Synthetic RNA-protein modules integrated with native translation mechanisms to control gene expression in malaria parasites. Nature Commun. 2016;7:10727. doi: 10.1038/ncomms10727. PMID: 26925876; PMCID: PMC4773503.

52. Wu W, Herrera Z, Ebert D, Baska K, Cho SH, DeRisi JL, et al. A chemical rescue screen identifies a *Plasmodium falciparum* apicoplast inhibitor targeting MEP isoprenoid precursor biosynthesis. Antimicrob Agents Chemother. 2015;59(1):356-64. doi: 10.1128/aac.03342-14. PMID: 25367906; PMCID: PMC4291372.

53. Mullin KA, Lim L, Ralph SA, Spurck TP, Handman E, McFadden GI. Membrane transporters in the relict plastid of malaria parasites. Proc Natl Acad Sci U S A. 2006;103(25):9572-7. doi: 10.1073/pnas.0602293103. PMID: 16760253; PMCID: PMC1480448.

54. Parsons M, Karnataki A, Feagin JE, DeRocher A. Protein trafficking to the apicoplast: deciphering the apicomplexan solution to secondary endosymbiosis. Eukaryot Cell. 2007;6(7):1081-8. doi: 10.1128/ec.00102-07. PMID: 17513565; PMCID: PMC1951102.

55. Calvo SE, Mootha VK. The mitochondrial proteome and human disease. Annu Rev Genomics Hum Genetics. 2010;11:25-44. doi: 10.1146/annurev-genom-082509-141720. PMID: 20690818; PMCID: PMC4397899.

56. Calvo SE, Clauser KR, Mootha VK. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. Nucleic Acids Res. 2016;44(D1):D1251-7. doi: 10.1093/nar/gkv1003. PMID: 26450961; PMCID: PMC4702768.

57. Boel G, Smith PC, Ning W, Englander MT, Chen B, Hashem Y, et al. The ABC-F protein EttA gates ribosome entry into the translation elongation cycle. Nat Struct Mol Biol. 2014;21(2):143-51. doi: 10.1038/nsmb.2740. PMID: 24389466; PMCID: PMC4101993.

58. Chen B, Boel G, Hashem Y, Ning W, Fei J, Wang C, et al. EttA regulates translation by binding the ribosomal E site and restricting ribosome-tRNA dynamics. Nat Struct Mol Biol. 2014;21(2):152-9. doi: 10.1038/nsmb.2741. PMID: 24389465; PMCID: PMC4143144.

59. Vazquez de Aldana CR, Marton MJ, Hinnebusch AG. GCN20, a novel ATP binding cassette protein, and GCN1 reside in a complex that mediates activation of the eIF-2 alpha kinase GCN2 in amino acid-starved cells. EMBO J. 1995;14(13):3184-99. PMID: 7621831; PMCID: PMC394380.

60. Marton MJ, Vazquez de Aldana CR, Qiu H, Chakraburtty K, Hinnebusch AG. Evidence that GCN1 and GCN20, translational regulators of *GCN4*, function on elongating ribosomes in activation of eIF2alpha kinase GCN2. Mol Cell Biol. 1997;17(8):4474-89. PMID: 9234705; PMCID: PMC232301.

61. Tyzack JK, Wang X, Belsham GJ, Proud CG. ABC50 interacts with eukaryotic initiation factor 2 and associates with the ribosome in an ATP-dependent manner. J Biol Chem. 2000;275(44):34131-9. doi: 10.1074/jbc.M002868200. PMID: 10931828.

62. Paytubi S, Wang X, Lam YW, Izquierdo L, Hunter MJ, Jan E, et al. ABC50 promotes translation initiation in mammalian cells. J Biol Chem. 2009;284(36):24061-73. doi: 10.1074/jbc.M109.031625. PMID: 19570978; PMCID: PMC2782000.

63. Sibley LD. The roles of intramembrane proteases in protozoan parasites. Biochim Biophys Acta. 2013;1828(12):2908-15. doi: 10.1016/j.bbamem.2013.04.017. PMID: 24099008; PMCID: PMC3793208.

64. Dowse TJ, Soldati D. Rhomboid-like proteins in Apicomplexa: phylogeny and nomenclature. Trends Parasitol. 2005;21(6):254-8. doi: 10.1016/j.pt.2005.04.009. PMID: 15922242.

65. Lin JW, Meireles P, Prudencio M, Engelmann S, Annoura T, Sajid M, et al. Loss-of-function analyses defines vital and redundant functions of the *Plasmodium* rhomboid protease family. Mol Microbiol. 2013;88(2):318-38. doi: 10.1111/mmi.12187. PMID: 23490234.

66. Lau JB, Stork S, Moog D, Schulz J, Maier UG. Protein-protein interactions indicate composition of a 480 kDa SELMA complex in the second outermost membrane of diatom complex plastids. Mol Microbiol. 2016;100(1):76-89. doi: 10.1111/mmi.13302. PMID: 26712034.

67. Sidik SM, Huet D, Ganesan SM, Huynh MH, Wang T, Nasamu AS, et al. A genome-wide CRISPR screen in *Toxoplasma* identifies essential apicomplexan genes. Cell. 2016;166(6):1423-35.e12. doi: 10.1016/j.cell.2016.08.019. PMID: 27594426; PMCID: PMC5017925.

68. Balabaskaran Nina P, Morrisey JM, Ganesan SM, Ke H, Pershing AM, Mather MW, et al. ATP synthase complex of *Plasmodium falciparum*: dimeric assembly in mitochondrial membranes and resistance to genetic disruption. J Biol Chem. 2011;286(48):41312-22. doi: 10.1074/jbc.M111.290973. PMID: 21984828; PMCID: PMC3308843.

69. Deitsch K, Driskill C, Wellems T. Transformation of malaria parasites by the spontaneous uptake and expression of DNA from human erythrocytes. Nucleic Acids Res. 2001;29(3):850-3. PMID: 11160909; PMCID: PMC30384.

70. Wagner JC, Platt RJ, Goldfless SJ, Zhang F, Niles JC. Efficient CRISPR-Cas9-mediated genome editing in *Plasmodium falciparum*. Nat Methods. 2014;11(9):915-8. doi: 10.1038/nmeth.3063. PMID: 25108687.

71. El Bakkouri M, Pow A, Mulichak A, Cheung KL, Artz JD, Amani M, et al. The Clp chaperones and proteases of the human malaria parasite *Plasmodium falciparum*. J Mol Biol. 2010;404(3):456-77. doi: 10.1016/j.jmb.2010.09.051. PMID: 20887733.

72. Tonkin CJ, van Dooren GG, Spurck TP, Struck NS, Good RT, Handman E, et al. Localization of organellar proteins in *Plasmodium falciparum* using a novel set of transfection vectors and a new immunofluorescence fixation method. Mol Biochem Parasitol. 2004;137(1):13-21. doi: 10.1016/j.molbiopara.2004.05.009. PMID: 15279947.

73. Gallagher JR, Prigge ST. *Plasmodium falciparum* acyl carrier protein crystal structures in disulfide-linked and reduced states and their prevalence during blood stage growth. Proteins. 2010;78(3):575-88. doi: 10.1002/prot.22582. PMID: 19768685; PMCID: PMC2805782.

74. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 2007;4(3):207-14. doi: 10.1038/nmeth1019. PMID: 17327847.

75. Huttlin EL, Jedrychowski MP, Elias JE, Goswami T, Rad R, Beausoleil SA, et al. A tissue-specific atlas of mouse protein phosphorylation and expression. Cell. 2010;143(7):1174-89. doi: 10.1016/j.cell.2010.12.001. PMID: 21183079; PMCID: PMC3035969.

76. Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol. 2004;340(4):783-95. doi: 10.1016/j.jmb.2004.05.028. PMID: 15223320.