A prioritized and validated resource of mitochondrial proteins in *Plasmodium* identifies leads to unique biology

Selma L. van Esveld[1,2], Lisette Meerstein-Kessel[1,3,‡], Cas Boshoven[4,‡], Jochem F. Baaij[1], Konstantin Barylyuk[5], Jordy P. M. Coolen[4], Ronald A.J. Duim[1], Bas E. Dutilh[1,6], Daniel R. Garza[1,7], Marijn Letterie[4], Nicholas I. Proellochs[4], Michelle N. de Ridder[4], Prashanna Balaji Venkatasubramanian[1], Laura E. de Vries[4], Ross F. Waller[5], Taco W.A. Kooij[4], Martijn A. Huynen[1,2]

[1]*Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboudumc, Nijmegen, the Netherlands*
[2]*Radboud Center for Mitochondrial Medicine, Radboudumc, Nijmegen, the Netherlands*
[3]*Radboud Institute for Health Sciences, Radboudumc, Nijmegen, the Netherlands.*
[4]*Department of Medical Microbiology, Radboudumc Center for Infectious Diseases, Radboud Institute for Molecular Life Sciences, Radboudumc, Nijmegen, the Netherlands*
[5]*Department of Biochemistry, University of Cambridge, Cambridge, UK*
[6]*Theoretical Biology and Bioinformatics, Science for Life, Utrecht University, Utrecht, the Netherlands*
[7]*Laboratory of Molecular Bacteriology (Rega Institute), Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium*

[‡]*These authors contributed equally to this work and should be considered co-second authors.*

Abstract

Plasmodium species have a single mitochondrion that is essential for their survival and has been successfully targeted by anti-malarial drugs. Most proteins are imported into this organelle and our picture of the *Plasmodium* mitochondrial proteome remains incomplete. Many data sources contain information about mitochondrial localization, including proteome and gene expression profiles, orthology to mitochondrial proteins from other species, co-evolutionary relationships, and amino acid sequences, each with different coverage and reliability. To obtain a comprehensive, prioritized list of *Plasmodium falciparum* mitochondrial proteins, we rigorously analyzed and integrated eight datasets using Bayesian statistics into a predictive score per protein for mitochondrial localization. At a corrected false discovery rate of 25%, we identified 295 proteins with a sensitivity of 65% and a specificity of 98%. They include proteins that have not been identified as mitochondrial in other eukaryotes but have characterized homologs in bacteria that are involved in metabolism or translation. Mitochondrial localization of seven *Plasmodium berghei* orthologs was confirmed by epitope labeling and co-localization with a mitochondrial marker protein. One of these belongs to a newly identified apicomplexan mitochondrial protein family that in *P. falciparum* has four members. With the experimentally validated mitochondrial proteins and the complete ranked *P. falciparum* proteome, which we have named PlasmoMitoCarta, we present a resource to study unique proteins of *Plasmodium* mitochondria.

Introduction

Although all mitochondria evolved from a single endosymbiotic event (1), they display a large variety among the eukaryotes. For example, while in most species the mitochondrial organelle still retains its genome, the number of proteins encoded varies from 100 in *Andalucia godoyi* (2) to three in myzozoan species like *Plasmodium falciparum* that only encode cytochrome b, cytochrome c oxidase subunits 1 and 3, and two fragmented rRNAs (3). Moreover, mitochondria-like organelles of a variety of anaerobic species have lost the

organellar genome altogether (4). Also, mitochondrial proteomes vary broadly in size, ranging from e.g. ~1500 different proteins in mammalian mitochondria (5) to a few members of only a single biochemical pathway found in some mitosomes (6). Within this variety, the *P. falciparum* mitochondrion, based on the size of its genome, the size of the nuclear genome, and the complexity of its oxidative phosphorylation, is predicted to have a relatively small proteome. Although several mitochondrial pathways have already been identified in *P. falciparum* (7), and databases like PlasmoDB contain valuable information about mitochondrial localization of individual proteins (8), a complete resource that weighs and combines all the relevant information about possible mitochondrial localization is lacking.

Plasmodium mitochondria are particularly interesting because their function is variable between different life-cycle stages. In the sexual blood stage, functional oxidative phosphorylation is essential for colonization and development inside the mosquito hosts (9, 10), while in asexual blood stages the only essential function of the respiratory chain is to recycle ubiquinone for pyrimidine biosynthesis (11). This is further highlighted by the observation that in asexual blood-stage parasites no cristae are detectable in the mitochondrial membrane, while in gametocyte mitochondria cristae-like structures that typically accumulate respiratory chain complexes are observed (10). A recent complexome profiling study demonstrated that the level of complex V components in asexual blood-stage parasites was only 3% of the level in gametocytes (12). Another intriguing aspect of mitochondria in *Plasmodium* and closely related species is that they are the only mitochondria that have no detectable homolog of the FeS cluster assembly protein frataxin (13), a protein that occurs even in mitosomes (14).

Knowledge on the mitochondrial proteome could provide viable targets for the development of new drugs against this deadly parasite, as apicomplexan mitochondrial proteomes contain unique proteins that are not present in the animal hosts (15). Recently for example, an unusual prohibitin-like protein, PHBL, was identified in *Plasmodium berghei* that might provide a transmission-blocking drug target as it is unique to Myzozoa and *phbl*$^{-}$ parasites fail to transmit (16). Furthermore, respiratory protein complex III has proven to be a good drug target (17) and inhibitors to the mitochondrial enzymes dihydroorotate dehydrogenase (DHODH) and NADH type II oxidoreductase (PfNDH2) have been identified (18).

For years, it has been difficult to perform proteomics experiments to identify the *Plasmodium* mitochondrial proteome due to the lack of organelle isolation methods that reliably separate mitochondria from the apicoplast (19). The definition of mitochondrial functions relied heavily upon homology studies (7, 20) and microscopic examination of individual proteins (Supplemental Table 1). With the use of two biotin tagging approaches, 422 mitochondrial matrix proteins were identified in *Toxoplasma gondii*, a species related to *Plasmodium* that also has an apicoplast and a mitochondrion (15). Besides these mitochondria-targeted approaches, hyperLOPIT, a whole-cell biochemical fractionation technique, defined the subcellular localization of the *T. gondii* proteome. This provided among others a set of 220 soluble and 168 membrane-bound *T. gondii* mitochondrial proteins (21). Alongside these proteomic data, there is a wealth of other data sources providing information on mitochondrial localization in *P. falciparum*. The challenge remains to use and combine these heterogeneous datasets to reliably predict the mitochondrial proteome.

Here, we integrated eight datasets containing proteomic, gene expression, orthology to mitochondrial proteins from other species, phylogenetic distribution, and

amino acid sequence data in a Bayesian manner as has been applied before such as for the definition of the mitochondrial proteome of humans (5, 22-25). We used a naïve Bayesian classifier to combine the data. This method exploits the relative strengths of the various datasets while maintaining transparency about the contribution of each dataset to the final Bayesian posterior probability score for being mitochondrial. For each protein, the contribution of each dataset to the prediction was determined using gold standards of *Plasmodium* proteins that are either known to be mitochondrial or non-mitochondrial. After data integration, we obtained a list of 295 P. falciparum mitochondrial proteins. We experimentally validated these predictions with seven proteins of varying probability scores, and collectively this predicted proteome indicates that the P. falciparum mitochondria are unique relative to mitochondria of model organisms.

Materials and Methods

Animal experiments. All animal experiments were performed in accordance with the Dutch Experiments on Animals Act (Wod), the Directive 2010/63/EU from the European Union and the European ETS 123 convention and approved by the Radboud University Animal Welfare Body (IvD) and Animal Experiment Committee (RUDEC; 2015-0142), and the Central Authority for Scientific Procedures on Animals (CCD; AVD103002016424). In this study, we used outbred male and female NMRI mice (Envigo).

Nuclear encoded reference proteome. All datasets used in this study were mapped to the *P. falciparum* 3D7 reference proteome version 3.1 from the Sanger Institute, downloaded December 2017. This version, including isoforms, contains 5431 proteins encoded by 5357 genes. As we are interested in the proteins encoded by nuclear genes, all apicoplast and mitochondrial encoded proteins were excluded, leaving a reference of 5324 genes. When a dataset contains informational data points for two or more protein isoforms of one gene, the value chosen for that gene was the one that comes closest to the expected values for mitochondrial proteins in that specific dataset.

Assembly of positive and negative lists for benchmarking. Bayesian data integration depends on gold standards of proteins known to be mitochondrial or non-mitochondrial. We constructed two gold standards to assess the predictive value of the individual datasets. Furthermore, we built two alternative sets for the construction of the CLIME (co-evolution) and WICCA (co-expression) datasets to prevent that the standards chosen to train those two predictors were also used to evaluate them. These four sets are unique, have no overlapping genes, and are available in Supplemental Table 1. To construct these lists, we commenced with a systematic review of all literature available via PubMed (as per 30-08-2018) using the following broad search string "(mitochondri* OR apicoplast OR plastid) AND (plasmodium OR malaria)". Given the intimate functional and physical relation between mitochondrion and apicoplast, we reasoned that the positive gold standard list should only include genes encoding proteins that had been unambiguously and singularly assigned to the mitochondrion through fluorescence microscopy and co-localization with confirmed mitochondrial markers or via immune-electron microscopy (54 genes). Any proteins for which multiple studies showed different results were excluded. To facilitate discrimination from the apicoplast, we also made a gold standard apicoplast list (72 genes). Finally, all proteins that were unambiguously non-mitochondrial as demonstrated in the same papers (170 genes) were combined with the apicoplast-positive

147 list to form the basis for the negative gold standard. This negative set was then
148 complemented with an additional 346 non-overlapping genes encoding non-apicoplast
149 and non-mitochondrial proteins identified through extensive literature review and
150 published online by the Ralph lab (26). The 588 genes encoding non-mitochondrial
151 proteins were randomly assigned to the negative gold standard (294 genes) that was used
152 for the Bayesian data integration and the alternative negative gold standard (294 genes)
153 that was used to train the CLIME and WICCA approaches. For the latter purpose, an
154 alternative positive standard was compiled of 146 genes, non-overlapping with the first
155 gold standard, but associated with one or more mitochondrial GO-terms (GO:0000275,
156 GO:0005739-43, GO:0005746, GO:0005750, GO:0005753, GO:0005758, GO:0005759,
157 GO:0006122, GO:0006839, GO:0006850, GO:0031966, GO:0033108, GO:0042775,
158 GO:0044429, GO:0044455).
159
160 **Datasets to predict mitochondrial and non-mitochondrial proteins.** To predict
161 mitochondrial proteins and non-mitochondrial proteins, we collected and constructed
162 eight datasets (Table 1) that contain features typical of either mitochondrial or apicoplast
163 proteins. Some proteins, like lipoate protein ligase 2 (PF3D7_0923600) (27) and serine
164 hydroxymethyltransferase isoforms (PF3D7_1235600) (28), are dual localized to the
165 mitochondrion and apicoplast, highlighting the need for the second category of datasets.
166 Each dataset is described in detail below, including how the dataset was processed to
167 obtain per *P. falciparum* protein a single score relevant to whether the protein is
168 mitochondrial or not.
169
170
171 **Table 1: Input datasets for Bayesian integration.**

Dataset	Abbreviations used in figures	Size of dataset	Fraction identified (%)	Fraction gold identified (%)	P - value
Dataset based on microarray evidence					
Co-expression, built using in-house developed WICCA tool (rank ≤ 2400) (https://wicca.cmbi.umcn.nl/)	Co-expression	5310	45.6	83.0	2.0E-08
Datasets based on ortholog evidence					
hyperLOPIT (TAGM.MCMC.joint.mitochondrion_max ≥ 4.78E-05) (21)	hyperLOPIT	2056	16	94.9	3.3E-28
Mitochondrial ortholog	Mito ortholog	5280	8.7	75.5	2.6E-32

Evolutionary inference, built using CLIME (NN score ≤ -1.5) (29)	CLIME	4687	9.0	38.5	6.9E-09
Absence of a *Cryptosporidium* ortholog (score ≤ 4)	Crypto ortholog	5280	67.5	77.4	7.9E-02
Datasets based on amino acid sequence evidence					
Mitochondrial targeting signal, built using PFMpred tool (SVM ≥ -0.67) (30)	Mitochondrial TS	5324	59.5	83.3	1.4E-04
Absence of an Apicoplast Targeting Signal, based on PlasmoAP (31)	Apicoplast TS	5324	90.5	96.3	1.0E-01
Isoelectric point (Patrickios pI ≥ 4.2)	pI	5324	55.1	83.3	1.1E-05

The size of datasets indicates the number of *Plasmodium* genes that could potentially have been measured in this dataset. Fractions indicate the actually identified genes of this set that fall in bins with a positive mitochondrial score (Fig 1A). P values (Fisher's exact test) indicate overrepresentation of identified gold standard genes compared to random expectation.

Dataset based on microarray evidence

Co-expression. Co-expressed genes tend to code for proteins that functionally interact. For mammals, co-expression analyses have been successfully used in the discovery of mitochondrial proteins (5, 32). To determine if the expression pattern of a *Plasmodium* gene correlates with mitochondrial protein coding genes, we developed a co-expression tool: WeIghted Co-expression Calculation tool for plAsmodium genes (WICCA, available at https://wicca.cmbi.umcn.nl/). In short, the tool uses the co-expression of a group of input genes with each other to weigh 83 microarray experiments for their predictive value for that group. Datasets that show high co-expression of the input genes are more likely to be relevant for the system that the input genes are part of, and receive a higher weight. It then uses these weights to combine the expression datasets and calculate one co-expression score per gene with the input system, in this case, one score per gene for the level of co-expression with mitochondrial protein coding genes. We used the co-expression ranking obtained with the alternative positive standard of 146 genes as input for WICCA to weigh the co-expression datasets. The methods of this tool were based on the WeGET method (http://weget.cmbi.umcn.nl/, (33)), methodological details on WICCA can be found in the supplement.

Datasets based on ortholog evidence

hyperLOPIT. Hyperplexed Localization of Organelle Proteins by Isotope Tagging (hyperLOPIT) is a proteomics technique that analyzes protein distributions upon biochemical fractionation. It enables the identification of the subcellular localization of

200 thousands of proteins (34, 35). Barylyuk *et al.* used this technique on the apicomplexan
201 *T. gondii* and, among others, classified proteins identified in all three hyperLOPIT
202 experiments to be part of the mitochondrial soluble matrix and mitochondrial membrane.
203 This classification was based on t-augmented Gaussian mixture models (TAGM) in
204 combination with maximum a posteriori prediction (TAGM-MAP) and Markov-chain
205 Monte-Carlo (TAGM-MCMC) methods (21). Marker proteins used for the classification were
206 arbitrarily set to 1 in the class that they belong to and to 0 for all other classes. Using
207 BLAST (36), or if that produced no homologs HHpred (37), in combination with selecting
208 for best bidirectional hits, we determined a list of orthologs between *T. gondii* and
209 *P. falciparum*. The 3,832 *T. gondii* proteins that got a TAGM-MCMC probability for
210 mitochondrial soluble matrix and mitochondrial membrane were mapped to the
211 corresponding *P. falciparum* ortholog when available. This resulted in a list of 2,057
212 *P. falciparum* identifiers with two TAGM-MCMC probabilities, one for mitochondrial matrix
213 localization and one for mitochondrial membrane localization. As the input datasets for
214 the Bayesian integration should be independent, both TAGM-MCMC probabilities cannot
215 be included as separate datasets. Therefore the maximum TAGM-MCMC probability (so
216 either the soluble or the membrane value) for each protein was used to create one input
217 hyperLOPIT dataset.
218 **Mitochondrial ortholog.** We also used orthologs from more distantly related
219 species than *T. gondii*, as also at larger phylogenetic distances subcellular localizations
220 tend to be conserved (38). Using BLAST (36) (all versus all blast for proteomes with e-value
221 of 100) in combination with OrthoMCL (39) (percent match cut of 50, e-value exponent cut
222 off -5) to detect orthologs, or if that produced no result, HHpred (37) (default settings, e-
223 value cut off 0.01, three iterations, only best-bi-directional hits included), one-to-one
224 orthologs between *P. falciparum* and either *Homo sapiens*, *Arabidopsis thaliana* or
225 *Saccharomyces cerevisiae* were determined. When the orthologous protein of at least one
226 species was annotated to be mitochondrial in a published mitochondrial compendium (*H.*
227 *sapiens* MitoCarta2.0 (5), *A. thaliana* (40), *S. cerevisiae* (41)), then that *P. falciparum* protein
228 was included in the mitochondrial ortholog dataset. *P. falciparum* proteins that had
229 orthologs in plastids that were more similar than their orthologs in mitochondria were
230 excluded to prevent e.g. the apicoplast *P. falciparum* pyruvate dehydrogenase (PDH) that
231 is closely related to the *A. thaliana* plastid PDH to be included in the mitochondrial
232 orthologs.
233 **Evolutionary inference.** Clustered by inferred models of evolution (CLIME) uses
234 homology across model species (138 eukaryotic species and 1 prokaryotic outgroup) to
235 identify evolutionary conserved clusters. We downloaded the pre-computed CLIME
236 analyses of *P. falciparum* genes from http://gene-clime.org/ (29) and mapped the IDs to
237 our reference proteome. We used the CLIME matrix listing the presence/absence of
238 orthologs of *P. falciparum* in other species, with the alternative positive and negative sets
239 as training data, to train a perceptron with two hidden layers to obtain a single score per
240 protein (details in supplement). The resulting output scores on our test set formed the
241 CLIME input for the Bayesian integration.
242 **Absence of a *Cryptosporidium* ortholog.** Like *Plasmodium*, the genus
243 *Cryptosporidium* belongs to the phylum of Apicomplexa. *Cryptosporidium* species lack an
244 apicoplast and contain only a remnant mitochondrial-like organelle (42). Therefore it can
245 be expected that apicoplast proteins and many mitochondrial proteins will not have an
246 ortholog in *Cryptosporidium*. MetaPhOrs is an online tool that contains orthology and
247 paralogy predictions obtained from multiple phylogenetic trees (43) and was used to

248 assess orthology for *P. falciparum* genes in three *Cryptosporidium* species (*hominis,*
249 *parvum*, and *muris*). We calculated a combined score by multiplying two MetaPhOrs
250 metrics, one for the confidence level (see below) and one for the number of hits between
251 the four species, and used this combined score as input for the Bayesian integration. In
252 detail, the MetaPhOrs results include a consistency score (ranging from 0 with no overlap
253 to 1 if all trees contain the same protein relationship/orthology information) and an
254 evidence level for the number of independent databases (the theoretical maximum is 13,
255 but for our species combination it was four). The multiplication of consistency and
256 evidence level results in an arbitrary score (range 0-4) that indicates the confidence in
257 calling two proteins orthologs. *P. falciparum* proteins with a high score are more likely to
258 have an ortholog in *Cryptosporidium*. Notice that there is some overlap with the
259 Evolutionary inference as that includes one *Cryptosporidium* species. Nevertheless,
260 *Cryptosporidium* species show quite some variation in the complexity of their mitosomes
261 and the overlap in the predictions is limited (see below).
262
263 *Datasets based on amino acid sequence evidence*
264 **Mitochondrial targeting signal.** The canonical mitochondrial import system requires an
265 amphipathic α-helical N-terminal targeting sequence. As *Plasmodium* has a distinct amino
266 acid usage pattern (44), a *P. falciparum* specific tool, PFMpred (30), was chosen to predict
267 mitochondrial localization based on the amino acid sequence. This is a support vector
268 machine (SVM) based tool and was used in split-amino-acid-composition mode (Matthews
269 correlation coefficient 0.73) that allows separate calculations for the N-terminus, C-
270 terminus, and remainder of the protein. The tool reports one SVM-score per protein and
271 indicates whether the protein is predicted to be mitochondrial or not. The SVM-scores per
272 gene are used as the mitochondrial targeting signal dataset.
273 **Absence of an apicoplast targeting signal.** As for mitochondrial proteins, the
274 majority of the apicoplast proteins are encoded by the nucleus and targeted to the
275 apicoplast by an N-terminal signal and a transit peptide. PlasmoAP (31) apicoplast
276 localization predictions based on amino acid sequences were downloaded from
277 PlasmoDB (8) in January 2018. This binary dataset was used in the prediction to reduce
278 apicoplast protein contamination among the top scoring proteins.
279 **Isoelectric point.** Nuclear encoded mitochondrial proteins need to cross the
280 negatively charged mitochondrial membrane and need to function properly inside the
281 mitochondrial environment that has a slightly higher pH compared to the cytosol (45). A
282 density plot (Supplemental Fig 1) by Patrikios isoelectric points (46) of the human
283 mitochondrial proteome (MitoCarta2.0 (5)) showed that mitochondrial proteins on
284 average have a higher pI than the remainder of the proteome. Isoelectric points for the
285 *P. falciparum* proteome were calculated using the Patrickios algorithm (47) and directly
286 used as pI scores.
287
288 **Bayesian integration to predict mitochondrial proteins.** Using the eight datasets
289 described above and the gold standard evaluation sets, a mitochondrial score for each
290 *P. falciparum* reference protein was calculated. This score is the log odds of the likelihood
291 that a protein is localized in the mitochondrion relative to it being localized somewhere
292 else in the cell. For categorical data (mitochondrial ortholog, apicoplast targeting signal),
293 the dataset is separated into bins that represent the two categories. Continuous data (the
294 other six datasets) were binned in a systematic way using a custom python script
295 (https://github.com/JordyCoolen/Binning) (settings: --bins 0 --score 1). In short, this script

optimizes for each dataset the distribution of the bins such that the sum of log ratio scores of all bins combined (see below) is close to zero, as this will result in the best separation of positive and negative bins. The number of bins was varied to a maximum of five bins to achieve the most optimal separation without expanding the number such that there are too few proteins per bin to obtain reliable estimates of the log odds of that bin.

The fractions of gold standard proteins per bin are determined. The log ratio of these fractions determines the score for all other proteins in that respective bin. The mitochondrial score is based on the sum of log ratios of the individual datasets and is calculated as follows:

$$mitochondrial\ score = log_2\left(\frac{P_{mitochondrial}}{P_{non-mitochondrial}}\right) + \sum_{i=1}^{n} log_2\left(\frac{P(data_i|mitochondrial)}{P(data_i|non-mitochondrial)}\right)$$

$$with\ \frac{P(data_i|mitochondrial)}{P(data_i|non-mitochondrial)} = \frac{mitochondrial_pos_i}{mitochondrial_neg_i}$$

where mitochondrial_pos and mitochondrial_neg are the fractions of the positive and negative gold standard genes in sample i respectively. If there were no gold standard genes found in a certain bin, that positive or negative gold standard fraction was set to 0.5/'total number of negative set genes in the complete dataset' to prevent division by zero and allow calculation of the log ratio. The O_{prior}, $log_2(P_{mitochondrial}/P_{non\text{-}mitochondrial})$, is based on the estimation that 536 proteins out of the 5357 protein coding genes (~10%) encode a mitochondrial protein. It does not affect the ranking of potential mitochondrial proteins.

To assess the performance of the integration, a false discovery rate (FDR) was calculated. As this FDR depends on gold standard genes and the ratio gold positives/negatives is not similar to the ratio of mitochondrial protein coding genes versus non-mitochondrial coding genes in the genome, the FDR was corrected (cFDR) using the following formula:

$$cFDR = \frac{1 - specificity}{1 - specificity + sensitivity \cdot O_{prior}}$$

Cross-validation. To assess the ability of the integrated predictor to discriminate known mitochondrial proteins from non-mitochondrial proteins, ten-fold cross-validation was performed. The gold standards (both negative and positive) were subsampled ten times, thereby creating ten sets of nine-tenth of the gold standard genes. Each gold standard gene was left out once in one of the ten sets. Data integration was performed with each of these ten sets and the ranks of the one-tenth left out gold standard genes were retrieved. The ten-fold cross-validated ROC curve was constructed based on those ranks.

Intraspecies homologs. The human mitochondrial proteome has expanded since the divergence from yeast mainly due to gene duplications that create mitochondrial paralogs (38). It is therefore interesting to see if the predicted *P. falciparum* mitochondrial proteins have intraspecies homologs, and whether we can identify mitochondrial gene families. With HHpred (37), using default settings and a 1E-5 E-value cut-off, the homologs within

336 the genome of *P. falciparum* were determined and included as a column in Supplemental
337 Table 2.
338
339 **Validation of candidate proteins.** To validate the predictions, we selected 14 genes with
340 unusual or novel mitochondrial functions within the top 295 of genes that fall within the
341 25% cFDR cut-off. Of these, 13 were selected based on maximum transcript levels during
342 asexual blood-stage development of *P. berghei* strain ANKA cl15cy1 that exceeded 100
343 FPKM (48). In addition, we selected one particularly interesting gene, PF3D7_1004900 that
344 had a lower transcript level. We used experimental genetics (49) to generate *P. berghei*
345 lines expressing 3xHA-tagged copies of the selected proteins and colocalized them with
346 an established mitochondrial marker (16) to determine their subcellular localization by
347 immunofluorescence microscopy. Details on plasmid construction (successful for 13 out of
348 14 targets, excluding PF3D7_1142800) and the generation of the lines can be found as
349 supplemental information. To assess colocalization, the freshly harvested transgenic
350 parasites were allowed to settle on a poly-L-lysine coated coverslip for 10 minutes and
351 then fixed with 4% EM-grade paraformaldehyde (Fisher Scientific) and 0.0075% EM-grade
352 glutaraldehyde (Sigma-Aldrich) in microtubule stabilizing buffer (MTSB, 10 mM MES, 150
353 mM NaCl, 5 mM EGTA, 5 mM glucose, 5 mM MgCl2 pH 6.9) for 20 minutes (50). Next,
354 parasites were permeabilized with 0.1% Triton X-100 in PBS for 10 minutes. Samples in
355 which we imaged mOrange and mCherry were hereafter stained with DAPI (1:300) for 15
356 minutes and mounted on a microscope slide with VectaShield (Vector laboratories). All
357 other samples were blocked in 3% FCS-PBS for 1 hour. Samples were incubated overnight
358 at 4°C with rat anti-HA (1:500, ROAHAHA Roche) and chicken anti-GFP (1:1000, Thermo
359 Fisher) antibodies and for 1 hour with 1:500 goat anti-rat Alexa fluor™ 594 (Invitrogen) and
360 goat anti-chicken Alexa fluor™ 488 (Thermo Fisher) conjugated antibodies at room
361 temperature. Nuclei were stained with DAPI (1:300) for 15 minutes and coverslips were
362 mounted on a microscope slide with VectaShield (Vector laboratories). Samples were
363 imaged with an SP8 confocal microscope (Leica, 63x oil lens) or LSM900 confocal
364 microscope (Zeiss, 63x oil lens). Images were processed minimally and similarly with Fiji
365 (51).
366
367 **Tools for data analysis.** Plots, statistics, and calculations were performed with the R
368 statistical package (52) and additional packages gplots (53), ggplot2 (54), ROCR (55), scales
369 (56) and reshape (57). The separation of the datasets into bins was performed using a
370 python script (see above).
371
372 **Results**
373 We integrated eight features of proteomic, gene expression, orthology and amino acid
374 composition data, to predict mitochondrial localization in *P. falciparum*. All the individual
375 features had a predictive value for mitochondrial localization (Fig 1A and 1B), including, as
376 expected, a negative score for the "*Cryptosporidium* ortholog" and "apicoplast targeting
377 signal" datasets (Supplemental Fig 2A). Notice that for the co-expression feature it is the
378 bin with the low-rank values that contains most mitochondrial proteins because this bin
379 contains the proteins with the highest co-expression values. The best performing input
380 datasets are the ones that translate mitochondrial localization between species: the
381 hyperLOPIT dataset and the mitochondrial ortholog dataset, followed by the
382 mitochondrial targeting signal and co-expression datasets (Fig 1B). Correlation analyses
383 showed that the features are largely independent off each other, allowing us to use naïve

384 Bayesian data integration. The highest correlations were observed between features that
385 examine the phylogenetic distribution of the proteins: CLIME and *Cryptosporidium*
386 ortholog datasets (Supplemental Fig 2B). By integrating all data, we achieved both a
387 better sensitivity and specificity in predicting mitochondrial proteins with an AUC of 0.958
388 than we achieved for individual data sets (Fig 1B). The exception to this is the hyperLOPIT
389 dataset that on its own performs best with an AUC of 0.960 (Fig 1B). However, this dataset
390 contains information on only 2,056 (39%) of the nuclear encoded *Plasmodium* proteins,
391 highlighting the value of additional datasets and their integration to predict a complete
392 mitochondrial proteome. The distribution of the log-odds ratios (Fig 1C) showed that the
393 mitochondrial score separated the positive and negative gold standards well from each
394 other and that there are additional proteins with a high score that were not part of either,
395 which are potential new mitochondrial proteins. We use a cFDR of 25% (Fig 2A) as a
396 threshold for highly probable new mitochondrial proteins and identify 295 proteins (Fig
397 2B) with a sensitivity of 65% and a specificity of 98%. The ranked list of the complete
398 nuclear encoded proteome is available in Supplemental Table 2.
399 Overall, the hyperLOPIT dataset predicts 388 *T. gondii* mitochondrial proteins with
400 a TAGM-MCMC location probability of 99% or higher. Of those, 290 have a *P. falciparum*
401 ortholog and of these orthologs, we find 217 in our predicted mitochondrial proteome.
402 We predicted 52 additional mitochondrial proteins that do not have a *T. gondii* ortholog
403 and 26 proteins that are present in the hyperLOPIT dataset but fell outside their 99% cut
404 off for mitochondrial localization.
405

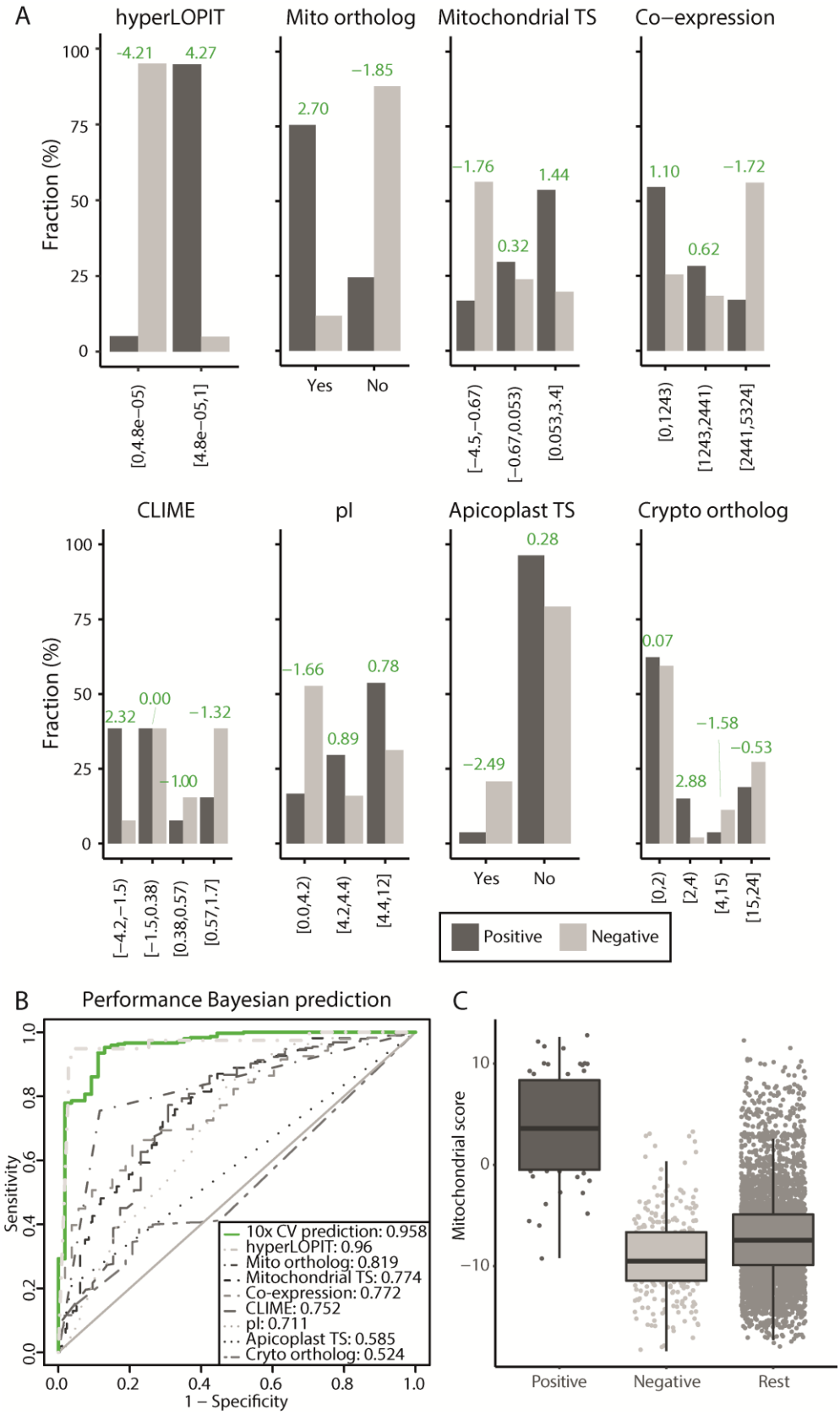

Fig 1. Predictive values of the individual datasets and the integrated ranked list. (A) Barplot of each dataset with the fraction of the gold standards on the y-axis and the bins on the x-axis. In green, the mitochondrial score per bin is indicated to show the predictive value of each bin. (B) ROC-curves comparing the performance to identify mitochondrial

proteins of the prediction (10-fold cross-validated) to the performance of the individual datasets. For comparison, the area under the curve values are indicated in the legend. (C) Boxplot that visualizes the distribution of the gold-standard positives, negatives, and remainder of the proteome (x-axis) over the calculated mitochondrial score of the prediction (y-axis).

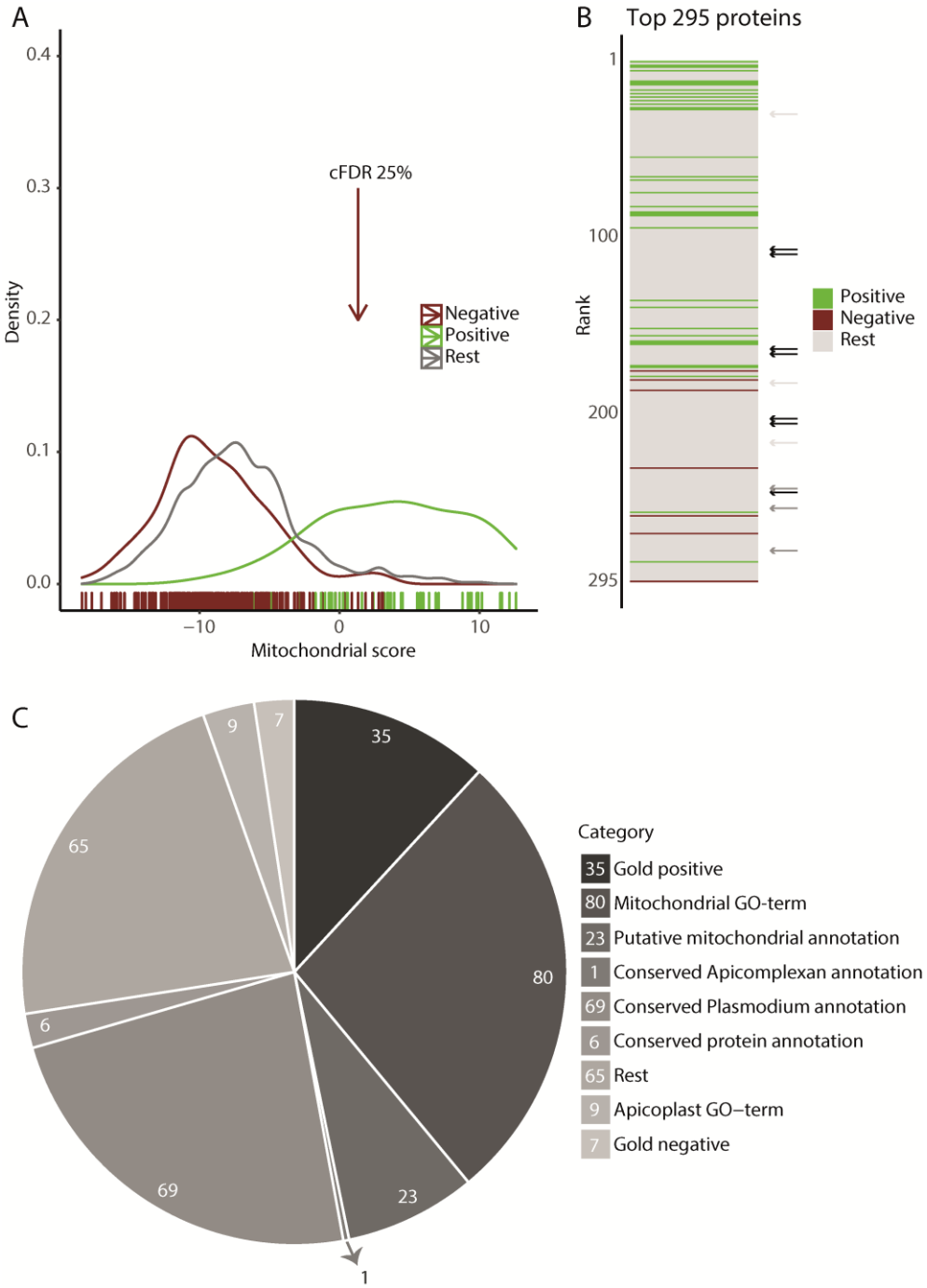

Fig 2. Integration produces a ranked proteome with 295 likely mitochondrial proteins at a 25% cFDR-cut. (A) Density plot of the mitochondrial score with a colored bar at the bottom indicating the scores of individual gold standards. The arrow indicates the 25% cFDR cut-off. (B) The top 295 proteins falling in the 25% cFDR cut-off ranked on their mitochondrial score with color indicating if the protein is part of a gold standard. Black arrows indicated the ranking of the seven candidate proteins that we experimentally

424 confirmed to be mitochondrial (Fig 3A and 3B), dark grey arrows indicate the ranking of
425 the three proteins with suggested mitochondrial localization (Fig 3C), and light grey
426 arrows the ranking of the three proteins with inconclusive results (Supplemental Fig 4B).
427 (C) Categorical representation of the 295 proteins identified at the 25% cFDR cut-off. First,
428 genes were separated into categories based on the gold standard, the remainder on
429 mitochondrial and apicoplast GO term, and subsequently the remainder on gene function
430 annotations.
431
432 **Validation of candidates.** We aimed to validate the mitochondrial candidate list by
433 tagging orthologs of the selected unusual or unique predicted mitochondrial proteins in
434 the efficient transfection model *P. berghei* (see the section 'new mitochondrial proteins' for
435 the *P. falciparum* orthologs of these proteins). Mitochondrial localization was assessed by
436 fluorescent microscopy using an experimental genetic approach developed in our lab
437 (16). This method allows the endogenous tagging of a target protein with a combined
438 fluorescent and epitope tag, while simultaneously introducing a strong mitochondrial
439 targeted GFP marker by fusion of the promoter and N terminus of HSP70-3
440 (PBANKA_0914400) to GFP. Initially, we tagged 6 proteins with a mOrange-3xHA tag. Live
441 imaging revealed very poor and undefined signals (data not shown). Next, we fixed blood
442 samples and stained them with anti-HA antibodies. Using this approach, only
443 PBANKA_0310100 and PBANKA_1203200 were convincingly shown to localize to the
444 mitochondrion (Fig 3A), while other proteins demonstrated mostly undetectable or
445 undefined signals (Supplemental Fig 4A). Since many of the selected proteins are
446 relatively small compared to the mOrange-tag and are expected to be imported into the
447 mitochondrion, we anticipated that interference of the tag with import or proper folding
448 could lead to the observed weak and undefined signals. To assess this, we selected two
449 similarly small and abundantly expressed targets, PBANKA_0310100 (127 amino acids) and
450 PBANKA_1024800 (144 amino acids), to test different tags. For PBANKA_0310100, which
451 was already successfully localized to the mitochondrion, these included tags consisting of
452 only the fluorescent protein mOrange or only the 3xHA epitope tag, either with or without
453 a linker sequence. For PBANKA_1024800, we also included the combined mCherry-cMyc
454 tag that has previously been used successfully to tag mitochondrial proteins (16). We
455 found that the fluorescent protein tag indeed interfered with the localization of
456 PBANKA_1024800, while the 3xHA tag, with or without a linker, supported a mitochondrial
457 localization (Supplemental Fig 5B). Surprisingly, using the mOrange-tag without the
458 subsequent HA-tag also led to an undefined rather punctuate mislocalization of
459 PBANKA_0310100 (Supplemental Fig 5B). Based on these observations, we decided to tag
460 the remaining 11 proteins with a linker-3xHA-tag (Fig 3B, 3C and Supplemental Fig 4B).
461 Five of these were also convincingly localized to the mitochondria (Fig 3B), while an
462 additional three proteins showed very weak staining patterns suggestive of a possible
463 mitochondrial localization (Fig 3C). The signals of the remaining three tagged proteins
464 were unable to be distinguished from the background and therefore no subcellular
465 localization could be assigned (Supplemental Fig 4B).
466

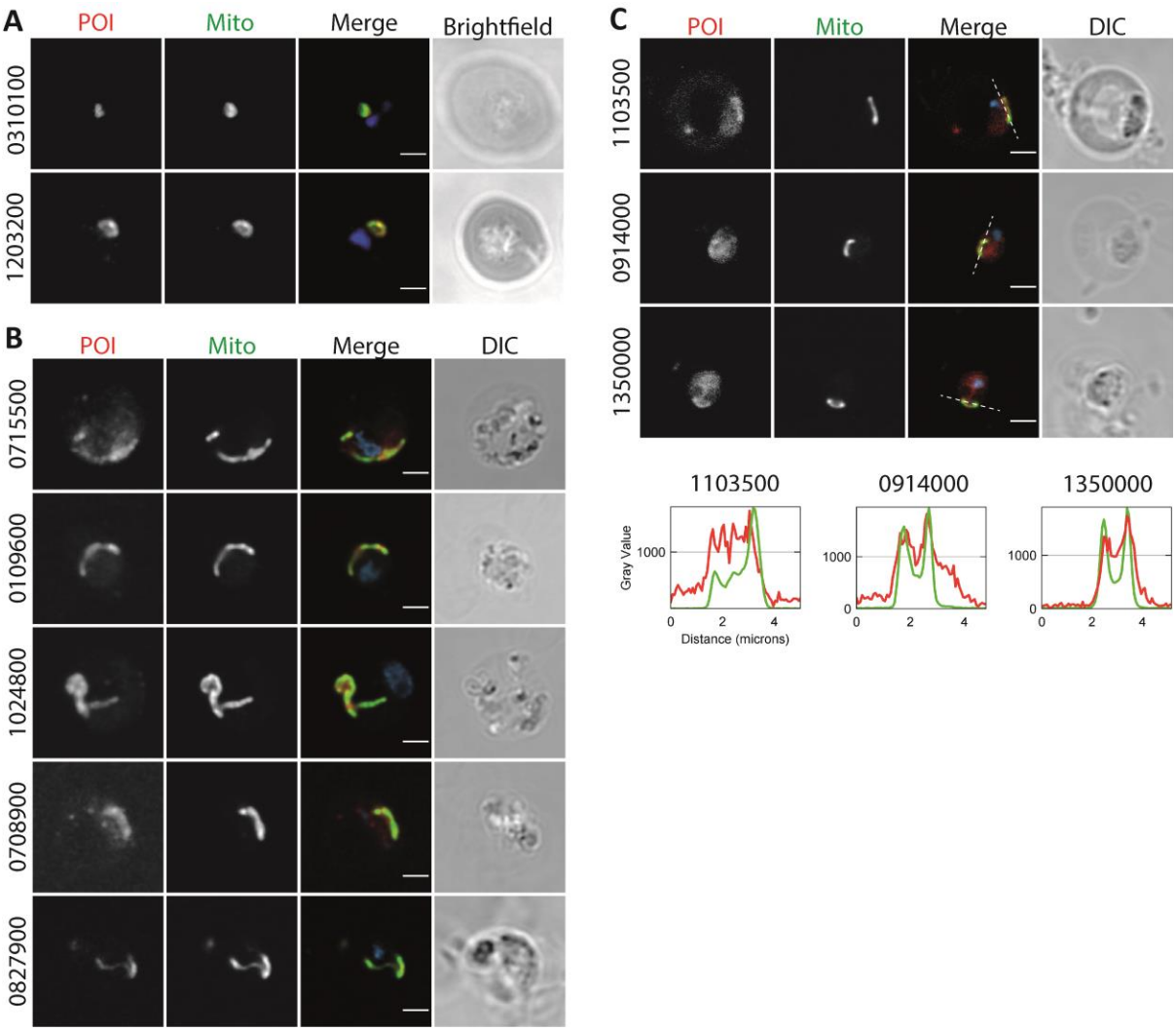

Fig 3. Validation of mitochondrial candidate proteins. Immunofluorescent analysis of tagged candidate proteins of interest (POI) in *P. berghei* parasites (PBANKA_ID indicated vertically). POI were stained with anti-HA antibody (red, first column), mitochondria were stained with anti-GFP antibody binding to the organelle marker (Mito, green, second column). The DNA was stained with DAPI (blue in merge). (A) Representative images of candidate proteins tagged with a mOrange-3xHA tag, imaged on a SP8 confocal microscope (Leica). (B-C) Representative images of candidate proteins tagged with a linker-3xHA tag, imaged on a LSM900 confocal microscope (Zeiss). (C) Images with the brightest signal for three lowly expressed candidate proteins. Pixel intensities of the POI and mitochondrial signal were normalized and plotted over the line indicated in the merge panel. Scale bar, 2 µm.

New mitochondrial proteins. A categorical representation of the identified proteins (Fig 2C) shows that 138 of the 295 proteins have a previous annotation of being mitochondrial as they are either part of the gold standard, have a mitochondrial GO term, or are described as putative mitochondrial protein. A smaller number of proteins, 16, were part of the negative set or have an apicoplast GO-term & no mitochondrial GO term. The remaining 141 proteins, of which 76 proteins are annotated as conserved hypothetical proteins, contain new mitochondrial candidates.

487 For 14 of the 76 conserved proteins with an unknown function, we found a
488 mitochondrial ortholog using the sensitive homology detection tool HHpred (37) that can
489 improve their annotation (Supplemental Table 2). For example, the human and yeast
490 orthologs of PF3D7_0306000 (Bayesian rank 41) are both cytochrome b-c1 complex
491 subunit 8, making it very likely that the *Plasmodium* protein is also cytochrome b-c1
492 complex subunit 8, as confirmed by recent complexome profiling (12). Other similar
493 examples include PF3D7_1357600 (rank 283) that is likely a mitochondrial member of the
494 39S ribosomal protein family L53, and PF3D7_1309900 (rank 78) that is likely succinate
495 dehydrogenase assembly factor 2.
496 Besides the possibility to improve annotations using well-characterized
497 mitochondrial orthologs, the predicted 295 proteins contain biologically interesting
498 proteins that can add new mitochondrial functions to *P. falciparum* and related species.
499 Therewith, the prediction provides targets for experimental research to uncover new
500 biology. These proteins can be divided into five categories: (i) proteins with known
501 mitochondrial orthologs and an unknown mitochondrial role in *P. falciparum*, (ii) proteins
502 without a mitochondrial ortholog but with homology to proteins with a known function
503 and conservation of critical residues that suggest conservation of that function, (iii)
504 proteins with homology to proteins with a known function from which residues known to
505 be essential for catalytic activity or for binding substrates residues have been lost, (iv) new
506 proteins of a known mitochondrial protein family and (v) proteins of a new mitochondrial
507 protein family. Below, we describe several proteins per category, including their domain
508 structure, the conservation of active site residues, and their phylogenetic distribution.
509
510 *Proteins with known mitochondrial orthologs and an unknown mitochondrial role in*
511 *P. falciparum*
512 PF3D7_0413500 (rank 109) contains a PGM domain and appears to be orthologous to the
513 human serine/threonine-protein phosphatase, PGAM5, of which the catalytic residue
514 H105 (58) is conserved. In human cells, PGAM5 is localized to the mitochondrial outer
515 membrane (59) where it plays a role in apoptosis, mitofission, and mitophagy (reviewed in
516 (60)) among others via interactions with BCL2-like protein 1 (61). We experimentally
517 confirmed PF3D7_0413500's mitochondrial location (PBANKA_0715500 in Fig 3B), making
518 it an interesting protein to study in the context of *Plasmodium* mitochondrial dynamics, a
519 process about which little is known. Originally, we flagged two additional proteins,
520 PF3D7_0213200 (rank 164) and PF3D7_0611300 (rank 166), as interesting due to potential
521 homology to BCL2, based on sequence similarity and a similar alpha-helical structure
522 predicted with HHpred (37). PF3D7_0213200 was assigned an apicoplast GO-term, but in
523 our analysis is a potential mitochondrial protein. Analysis with TMHMM (62) indicates that
524 it probably is a transmembrane protein with an in-out topology. The protein
525 PF3D7_0611300 already was assigned a mitochondrial GO-term based on a screening in
526 *T. gondii* (63), but was, to our knowledge, not confirmed as mitochondrial localized with
527 with experiments characterizing this protein in *P. falciparum.* Although we rejected BCL2
528 homology for both proteins after a detailed examination of the residues conserved in the
529 BCL2 family, we confirmed the mitochondrial localization of PF3D7_0213200
530 (PBANKA_0310100 in Fig 3A) and PF3D7_0611300 (PBANKA_0109600 in Fig 3B). Indeed,
531 complexome profiling revealed that the latter is part of the ATP synthase complex (12).
532
533 *Proteins with a non-mitochondrial homolog and conservation of critical residues*

534 PF3D7_0913400 (rank 30) is homologous to the bacterial elongation factor P. Although
535 eukaryotes contain initiation factor 5A as a cytoplasmic elongation factor P homolog, this
536 protein family was until now not observed in mitochondria. PF3D7_0913400 contains
537 both the N-terminal KOW-like domain and the central P/YeiP domain of elongation factor
538 P, but not the C-terminal OB domain. Elongation factor P plays a role in translation
539 elongation and specifically in the translation of stretches of 2 or more prolines (64). A basic
540 residue at position 34 of elongation factor P (numbering from the *Escherichia coli* PDB
541 structure 3A5Z_F (65)) that after post-translational modification interacts with the CCA-
542 end of the tRNA, is conserved in PF3D7_0913400. Notably, mitochondrial encoded
543 *P. falciparum* cytochrome C oxidase subunit 1 contains a proline pair at position 135-136,
544 making it a potential target for elongation factor P.
545 PF3D7_0812200 (rank 144) is homologous to the bacterial DegP protease family. It
546 contains both a serine endoprotease domain and a PDZ domain. The DegP from *E. coli*
547 functions in acid resistance in the periplasm. It is able to refold after acid stress and
548 subsequently can cleave proteins misfolded due to acid stress (66). The active site residue
549 (S210 in PDB structure 1KY9 (67)) is conserved in PF3D7_0812200, but residues involved in
550 allosteric loop interactions, including R187 are not. Orthologs of PF3D7_0812200 appear,
551 among the eukaryotes, limited to the Apicomplexa and Chromerida.
552 PF3D7_0503900 (rank 242) has a C-terminal dioxygenase domain (E= 1.3E-16). The
553 closest homolog in human is the phytanoyl-CoA dioxygenase that resides in the
554 peroxisome, an organelle absent from *Plasmodium*, where it is involved in α-oxidation of
555 branched-chain fatty acids. Important residues in the active site that are involved in iron
556 binding, H175, H177, H264 (68), are conserved in PF3D7_0503900, suggesting that it is a
557 functional enzyme. Nevertheless, levels of sequence conservation are low (14% identity
558 with the human phytanoyl-CoA dioxygenase) and we did not detect conservation of
559 substrate binding sites with the phytanoyl-CoA dioxygenase. The co-localization analysis
560 suggested a mitochondrial localization for PF3D7_0503900 (PBANKA_1103500 in Fig 3C).
561 PF3D7_1121500 (rank 250) contains a papain-like NLpC/P60 superfamily domain
562 (Peptidase_C92 in PFAM). This protein family contains, among others, cysteine peptidases
563 and amidases. Of the four residues essential for activity in BcPPNE from *Bacillus cereus* (69),
564 the most similar experimentally characterized homolog of PF3D7_1121500, three are
565 conserved (H49, E64, and Y164).
566
567 *Proteins with homology to proteins with a known function from which residues known to be*
568 *essential for catalytic activity or for binding substrates residues have been lost*
569 PF3D7_1004900 (rank 206) is homologous to the protein component of the signal
570 recognition particle (SRP), a ribonucleoprotein that recognizes and targets specific
571 proteins to the ER in eukaryotes. PF3D7_1004900 contains the C-terminal M domain (E=
572 5.8E-25), but not the other, SRP54 and SRP54_N domains of this protein. In SRP the M
573 domain binds both the SRP RNA and the signal sequence of the target protein. Inspection
574 of individual amino acids did not reveal conservation of RNA binding amino acids, e.g. the
575 arginines in helix M4 of the M domain (70), are not conserved, but did reveal some
576 conservation of hydrophobic amino acids lining the groove in which the signal peptide is
577 located. These hydrophobic amino acids have been implicated in interactions with the
578 signal peptide, specifically V323, I326, L329 in the M1 helix and L418 in the M5 helix
579 (positions relative to *Sulfolobus solfataricus* structure 3KL4 (70, 71)), suggesting a possible
580 conservation of interaction of PF3D7_1004900 with a hydrophobic α-helix. We

581 experimentally confirmed the mitochondrial location of PF3D7_1004900
582 (PBANKA_1203200 in Fig 3A).
583 PF3D7_1246700 (rank 217) is a member of the pyridoxamine 5′-phosphate oxidase
584 (PNPOx) like protein family that, besides pyridoxamine 5′-phosphate oxidase, also
585 contains the general stress protein 26 family. The most similar experimentally
586 characterized protein is general stress protein 26 from *Xanthomonas citri* encoded by gene
587 *Xac2369*. The latter binds FAD and FMN but not pyridoxal 5′-phosphate, indicating that it
588 does not function as a PNPOx (72). Relative to the unpublished structure of a protein from
589 this family from *Nostoc punctiforme* (PDB: 2I02), that does contain an FMN, we did not
590 observe conservation of the amino acids like W111, that line the FMN binding pocket,
591 casting doubt of a function of PF3D7_1246700 that includes the binding of FMN.
592
593 *New members of mitochondrial protein families*
594 PF3D7_1417900 (rank 107), PF3D7_1142800 (rank 167) and PF3D7_0927100 (rank 244)
595 potentially contain a CHCH domain. They all have two pairs of cysteines that are very well
596 conserved among their detectable homologs, and the nine amino acids between the
597 cysteines are predicted to form an α-helix (37). Nevertheless, as nothing else but the
598 cysteines appears to be conserved, the E-value with the PFAM CHCH domain entry, or with
599 any known mitochondrial protein, is not significant (E>0.01). Jackhmmer analysis (73)
600 detects homologs among the Apicomplexa, including a homolog in *Cryptosporidium*
601 *muris*, but not outside this taxon. Known proteins with a CHCH domain are localized to the
602 mitochondrial intermembrane space where they can participate in disulfide bond
603 formation, suggesting that PF3D7_1417900, PF3D7_1142800, and PF3D7_0927100 might
604 also be in the intermembrane space. Complexome profiling revealed that the first two are
605 part of the ATP synthase complex (12). We tested PF3D7_1417900 and PF3D7_0927100
606 experimentally and confirmed them to be mitochondrial (PBANKA_1024800 and
607 PBANKA_0827900 in Fig 3B respectively). Thus, based on our data and the complexome
608 profiling data all three proteins are confirmed to be mitochondrial, increasing their
609 likelihood that they indeed contain a CHCH domain.
610 *Plasmodium* contains a cytochrome C heme lyase (PF3D7_1224600, rank 126) and a
611 cytochrome C1 heme lyase (PF3D7_1203600, rank 251), which are both essential during
612 blood-stage development and have non-overlapping functions (74). We uncovered a third
613 homolog in this family at rank 183, PF3D7_1121200 (E=0.0096, HHpred (37)). Although the
614 homology covers domains II, III, and IV of heme lyases, of which domain II has been
615 implicated in binding Heme (75), residues that have been implicated in heme binding, like
616 H154 (coordinates for human protein (75)) and residues of domain I are not conserved in
617 PF3D7_1121200. The protein also occurs in ciliates and dinoflagellates but has no
618 orthologs outside of those taxa and therewith appears to be an alveolate-specific
619 duplication.
620
621 *A new Apicomplexan mitochondrial protein family*
622 We assessed to what extent *Plasmodium* mitochondrial proteins were part of families with
623 multiple mitochondrial members that resulted from duplications (Supplemental Table 2).
624 Such "intra compartmental protein duplications" have also increased the size of the
625 human mitochondrial proteome (38). Among the mitochondrial homologs, we uncovered
626 a new family in the Apicomplexa that in *Plasmodium* species has four representatives:
627 PF3D7_0821900 (rank 203), PF3D7_1336100 (rank 254), PF3D7_1134400 (rank 278), and
628 PF3D7_1228000 (rank 503). The family is characterized by a well-conserved WPP motif at

its N-terminus (Supplemental Fig 7), and we therefore name it the WPP family. Although PF3D7_1228000 does not rank sufficiently high to fall within the 25% cFDR cut-off, all proteins of the family are likely mitochondrial given their overall high ranking and the tendency of members of protein families to be localized in the same cellular location. For the protein family itself, we did not detect homologs outside the Apicomplexa and *Vitrella*. In several species including *Plasmodium coatneyi* (PCOAH_00054040), *Babesia bovis* (BBOV_III010100), and *Theileria orientalis* (MACL_00002472), the ortholog of PF3D7_1228000 is predicted to be fused with an rRNA pseudourydilate synthase hinting at a potential function of this family in rRNA maturation. Nevertheless, there are no experimental data supporting those gene fusions. We unequivocally confirmed the mitochondrial location of PF3D7_0821900 (PBANKA_0708900 in Fig 3B) and the co-localization analysis also suggested a mitochondrial localization for PF3D7_1336100 and PF3D7_1134400 (PBANKA_1350000 and PBANKA_0914000 in Fig 3C respectively). PF3D7_1228000 was excluded from the analysis due to low transcript levels.

Discussion

In this work, we ranked all nuclear encoded proteins of *P. falciparum* for their likelihood of being mitochondrially targeted and defined the top 295 proteins (cFDR 25%) as likely mitochondrial proteins (Supplemental Table 2). Despite the large variation between the datasets in the coverage of the nuclear encoded proteome and in the ability to identify mitochondrial proteins (Table 1, Fig 1), the probabilistic approach we used was able to integrate the data in an unbiased way by assessing the distribution of gold standard genes within each dataset during the scoring process.

Quality of the predictions. Several observations indicate the high quality of the predictions. First of all, the 10-fold cross-validation with an area under the curve of 0.958 shows the high predictive value and robustness of the prediction. Second, half of the predicted proteins have a previous annotation of being mitochondrial, even though these were not all used in weighing the datasets. Third, of thirteen proteins that we set out to validate experimentally, the seven for which we obtained a high enough expression to convincingly locate them in the cell, were all mitochondrial, while for three others the experimental data was consistent with a mitochondrial localization but equivocal due to low expression levels. These seven plus three were well distributed over the order of predicted new mitochondrial proteins (Fig 2B). Fourth, even though our top 295 contains seven proteins of the negative gold standard that was not manually supervised for all entries, manual examination of those seven showed that the majority are likely mitochondrial after all. Six of them were detected in the hyperLOPIT study (21) and of those five were annotated as mitochondrial (legend Supplemental Table 2). Finally, several proteins that we predicted to be mitochondrial have, since we made those predictions, been shown to be mitochondrial using other methods. Specifically, complexome profiling of mitochondria-enriched fractions has elucidated the composition of mitochondrial complexes II, III, IV, and V (12). We identify in our top 295 three members of complex II (of which two with unknown function), nine members of complex III (of which three with unknown function), ten members of complex IV (of which six with unknown function), and eight of complex V (all with unknown function). In general, most members of the protein complexes (described in (12)) were retrieved by our mitochondrial score (Supplementary Tables 2 and 5). Of the 24 electron transport chain members that were newly discovered by complexome profiling (12), 19 were in our top 295 (Supplemental Tables 2 and 5). The

677 results outlined above reinforce the robustness of the predictions and that the ranked list
678 of the complete nuclear encoded *P. falciparum* proteome, which we name
679 PlasmoMitoCarta, is a useful resource for *Plasmodium* researchers.
680
681 **New mitochondrial biology.** The results of the integration provide ample material for
682 exploring new mitochondrial biology as 101 out of the 295 proteins do not have a
683 detectable mitochondrial ortholog in yeast, human, or plant. Some, however, do have
684 homologs in bacteria that include conservation of critical residues, for example, the
685 elongation factor P whose function, if confirmed, would be unique to mitochondria. Also
686 outside the cutoff are interesting candidates for mitochondrial proteins, like
687 PF3D7_0213000 (rank 328) that has an ortholog in *T. gondii* whose mitochondrial
688 localization was confirmed microscopically (21) and PF3D7_1026200 (rank 332) that
689 contains an RNA binding THUMP domain and a UPF0020 RNA methyltransferase domain.
690 The most similar homolog of PF3D7_1026200 with experimentally determined function is
691 the bifunctional methyltransferase RlmKL from *E. coli*. PF3D7_1026200 is homologous to
692 the N-terminal methyltransferase domain of RlmKL (E=5E-38) that methylates G2445 in
693 the *E. coli* large subunit rRNA (76), and this nucleotide is conserved in the *P. falciparum*
694 mitochondrial rRNA (77).
695 As part of this study, we provide two other valuable resources. First, we included
696 the ortholog information gathered for this study that can improve the annotation of
697 *P. falciparum* proteins and provide researchers with additional hints for the function of a
698 protein (Supplemental Table 2). In addition, the WICCA tool that we developed for this
699 study is available online. It to allows any researcher to assess co-expression of *P. falciparum*
700 genes with a set of genes of four *Plasmodium* species across data from 83 microarray
701 expression experiments, by weighing the datasets for their relevance to the set of genes.
702 This can aid in the discovery of novel members of known pathways and molecular
703 systems.
704
705 **Size of the mitochondrial proteome.** We predict 295 mitochondrial proteins at the strict
706 cut off, set at cFDR 25%. When using the posterior probabilities (2^mitochondrial score) of
707 all *P. falciparum* genes to calculate the estimated number (E) of mitochondrial proteins by
708 summing them all up (E=sum($P_{posterior}$/($P_{posterior}$+1), see ref (78) for a detailed explanation),
709 we predicted the total size of the mitochondrial proteome to be 466 proteins. It therefore
710 appears that our prior odds of ~10%, which affects the total size calculation,
711 overestimated the number of mitochondrial proteins. When we lower the prior odds (see
712 ref (78)) to 8.7% (430 proteins) the number is consistent with the estimated size of the
713 proteome based on all data. In contrast, the presence of five likely mitochondrial proteins
714 in our gold negative sets leads to an underestimate of the size of the mitochondrial
715 proteome. We decided not to manually weed out such inconsistencies, as this cannot be
716 done systematically for all datasets, and applying this only to the apparent inconsistencies
717 would result in circular arguments.
718 The predicted total size of 466 proteins is relatively small compared to the
719 mitochondrial proteomes in human, plant, and yeast that consist of at least 900 proteins
720 (5, 40, 41). However, small mitochondrial proteomes have been reported before, for
721 example, 573 proteins in the distantly related ciliate *Tetrahymena thermophila* (79). In
722 agreement with our findings, a recent study on the mitochondrial proteome of *T. gondii*
723 that was not included in our integration reports a proteome of 421 proteins (15).
724

725 In conclusion, we combined the information of eight datasets with the curated gold
726 standards to rank all nuclear encoded *P. falciparum* proteins on their likelihood of being
727 mitochondrial in PlasmoMitoCarta. The value of this ranked proteome is shown by *in vivo*
728 validation of top scoring proteins in the closely related species *P. berghei* and will provide
729 an important resource for future investigations.
730
731 **Funding**
732 SLvE and CB were supported by PhD fellowships from the Radboud Institute for Molecular
733 Life Sciences, Radboudumc (Radboudumc JO ronde 2014 and #19-015a). NIP was
734 supported by a Marie-Slodowska Curie grant (790085), and TWAK and LEdV by the
735 Netherlands Organisation for Scientific Research (NWO-VIDI 864.13.009). BED was
736 supported by the Netherlands Organization for Scientific Research (NWO) Vidi grant
737 864.14.004 and European Research Council (ERC) Consolidator grant 865694: DiversiPHI.
738 KB was supported by a Leverhulme Trust and Isaac Newton Trust Fellowship (ECF-2015-
739 562), and RFW by the Wellcome Trust (214298/Z/18/Z).
740
741 **Competing interests**
742 The authors declare no competing interests.
743
744 **REFERENCES**
745
746 1. Gray MW, Burger G, Lang BF. Mitochondrial evolution. Science (New York, NY).
747 1999;283(5407):1476-81.
748 2. Burger G, Gray MW, Forget L, Lang BF. Strikingly bacteria-like and gene-rich
749 mitochondrial genomes throughout jakobid protists. Genome biology and evolution.
750 2013;5(2):418-38.
751 3. Wilson RJ, Williamson DH. Extrachromosomal DNA in the Apicomplexa. Microbiology
752 and molecular biology reviews : MMBR. 1997;61(1):1-16.
753 4. Makiuchi T, Nozaki T. Highly divergent mitochondrion-related organelles in
754 anaerobic parasitic protozoa. Biochimie. 2014;100:3-17.
755 5. Calvo SE, Clauser KR, Mootha VK. MitoCarta2.0: an updated inventory of mammalian
756 mitochondrial proteins. Nucleic acids research. 2016;44(D1):D1251-7.
757 6. Jedelský PL, Doležal P, Rada P, Pyrih J, Smíd O, Hrdý I, et al. The minimal proteome in
758 the reduced mitochondrion of the parasitic protist Giardia intestinalis. PloS one.
759 2011;6(2):e17285.
760 7. van Dooren GG, Stimmler LM, McFadden GI. Metabolic maps and functions of the
761 Plasmodium mitochondrion. FEMS microbiology reviews. 2006;30(4):596-630.
762 8. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, et al. PlasmoDB:
763 a functional genomic database for malaria parasites. Nucleic acids research.
764 2009;37(Database issue):D539-43.
765 9. Goodman CD, Siregar JE, Mollard V, Vega-Rodríguez J, Syafruddin D, Matsuoka H, et
766 al. Parasites resistant to the antimalarial atovaquone fail to transmit by mosquitoes. Science
767 (New York, NY). 2016;352(6283):349-53.
768 10. Sturm A, Mollard V, Cozijnsen A, Goodman CD, McFadden GI. Mitochondrial ATP
769 synthase is dispensable in blood-stage Plasmodium berghei rodent malaria but essential in
770 the mosquito phase. Proceedings of the National Academy of Sciences of the United States
771 of America. 2015;112(33):10216-23.

11. Painter HJ, Morrisey JM, Mather MW, Vaidya AB. Specific role of mitochondrial electron transport in blood-stage Plasmodium falciparum. Nature. 2007;446(7131):88-91.
12. Evers F, Cabrera-Orefice A, Elurbe DM, Lindert MK-t, Boltryk SD, Voss TS, et al. Composition and stage dynamics of mitochondrial complexes in Plasmodium falciparum. bioRxiv. 2020:2020.10.05.326496.
13. Pastore C, Adinolfi S, Huynen MA, Rybin V, Martin S, Mayer M, et al. YfhJ, a molecular adaptor in iron-sulfur cluster formation or a frataxin-like protein? Structure (London, England : 1993). 2006;14(5):857-67.
14. Goldberg AV, Molik S, Tsaousis AD, Neumann K, Kuhnke G, Delbac F, et al. Localization and functionality of microsporidian iron-sulphur cluster assembly proteins. Nature. 2008;452(7187):624-8.
15. Seidi A, Muellner-Wong LS, Rajendran E, Tjhin ET, Dagley LF, Aw VY, et al. Elucidating the mitochondrial proteome of Toxoplasma gondii reveals the presence of a divergent cytochrome c oxidase. eLife. 2018;7.
16. Matz JM, Goosmann C, Matuschewski K, Kooij TWA. An Unusual Prohibitin Regulates Malaria Parasite Mitochondrial Membrane Potential. Cell reports. 2018;23(3):756-67.
17. Fry M, Pudney M. Site of action of the antimalarial hydroxynaphthoquinone, 2-[trans-4-(4'-chlorophenyl) cyclohexyl]-3-hydroxy-1,4-naphthoquinone (566C80). Biochemical pharmacology. 1992;43(7):1545-53.
18. Stocks PA, Barton V, Antoine T, Biagini GA, Ward SA, O'Neill PM. Novel inhibitors of the Plasmodium falciparum electron transport chain. Parasitology. 2014;141(1):50-65.
19. Kobayashi T, Sato S, Takamiya S, Komaki-Yasuda K, Yano K, Hirata A, et al. Mitochondria and apicoplast of Plasmodium falciparum: behaviour on subcellular fractionation and the implication. Mitochondrion. 2007;7(1-2):125-32.
20. Seeber F, Limenitakis J, Soldati-Favre D. Apicomplexan mitochondrial metabolism: a story of gains, losses and retentions. Trends in parasitology. 2008;24(10):468-78.
21. Barylyuk K, Koreny L, Ke H, Butterworth S, Crook OM, Lassadi I, et al. A Comprehensive Subcellular Atlas of the Toxoplasma Proteome via hyperLOPIT Provides Spatial Context for Protein Functions. Cell Host Microbe. 2020;28(5):752-66.e9.
22. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science (New York, NY). 2003;302(5644):449-53.
23. van der Lee R, Feng Q, Langereis MA, Ter Horst R, Szklarczyk R, Netea MG, et al. Integrative Genomics-Based Discovery of Novel Regulators of the Innate Antiviral Response. PLoS computational biology. 2015;11(10):e1004553.
24. Meerstein-Kessel L, van der Lee R, Stone W, Lanke K, Baker DA, Alano P, et al. Probabilistic data integration identifies reliable gametocyte-specific proteins and transcripts in malaria parasites. Scientific reports. 2018;8(1):410.
25. van Esveld SL, Cansız-Arda Ş, Hensen F, van der Lee R, Huynen MA, Spelbrink JN. A Combined Mass Spectrometry and Data Integration Approach to Predict the Mitochondrial Poly(A) RNA Interacting Proteome. Frontiers in cell and developmental biology. 2019;7:283.
26. Woodcroft BJ, McMillan PJ, Dekiwadia C, Tilley L, Ralph SA. Determination of protein subcellular localization in apicomplexan parasites. Trends in parasitology. 2012;28(12):546-54.
27. Günther S, Wallace L, Patzewitz EM, McMillan PJ, Storm J, Wrenger C, et al. Apicoplast lipoic acid protein ligase B is not essential for Plasmodium falciparum. PLoS pathogens. 2007;3(12):e189.

819 28. Read M, Müller IB, Mitchell SL, Sims PF, Hyde JE. Dynamic subcellular localization of
820 isoforms of the folate pathway enzyme serine hydroxymethyltransferase (SHMT) through
821 the erythrocytic cycle of Plasmodium falciparum. Malaria journal. 2010;9:351.
822 29. Li Y, Calvo SE, Gutman R, Liu JS, Mootha VK. Expansion of biological pathways based
823 on evolutionary inference. Cell. 2014;158(1):213-25.
824 30. Verma R, Varshney GC, Raghava GP. Prediction of mitochondrial proteins of malaria
825 parasite using split amino acid composition and PSSM profile. Amino acids. 2010;39(1):101-
826 10.
827 31. Foth BJ, Ralph SA, Tonkin CJ, Struck NS, Fraunholz M, Roos DS, et al. Dissecting
828 apicoplast targeting in the malaria parasite Plasmodium falciparum. Science (New York, NY).
829 2003;299(5607):705-8.
830 32. Baughman JM, Nilsson R, Gohil VM, Arlow DH, Gauhar Z, Mootha VK. A
831 computational screen for regulators of oxidative phosphorylation implicates SLIRP in
832 mitochondrial RNA homeostasis. PLoS genetics. 2009;5(8):e1000590.
833 33. Szklarczyk R, Megchelenbrink W, Cizek P, Ledent M, Velemans G, Szklarczyk D, et al.
834 WeGET: predicting new genes for molecular systems by weighted co-expression. Nucleic
835 acids research. 2016;44(D1):D567-73.
836 34. Mulvey CM, Breckels LM, Geladaki A, Britovšek NK, Nightingale DJH, Christoforou A,
837 et al. Using hyperLOPIT to perform high-resolution mapping of the spatial proteome. Nature
838 protocols. 2017;12(6):1110-35.
839 35. Christoforou A, Mulvey CM, Breckels LM, Geladaki A, Hurrell T, Hayward PC, et al. A
840 draft map of the mouse pluripotent stem cell spatial proteome. Nature communications.
841 2016;7:8992.
842 36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search
843 tool. Journal of molecular biology. 1990;215(3):403-10.
844 37. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, et al. A Completely
845 Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. Journal of
846 molecular biology. 2018;430(15):2237-43.
847 38. Szklarczyk R, Huynen MA. Expansion of the human mitochondrial proteome by intra-
848 and inter-compartmental protein duplication. Genome biology. 2009;10(11):R135.
849 39. Li L, Stoeckert CJ, Jr., Roos DS. OrthoMCL: identification of ortholog groups for
850 eukaryotic genomes. Genome research. 2003;13(9):2178-89.
851 40. Rao RS, Salvato F, Thal B, Eubel H, Thelen JJ, Møller IM. The proteome of higher plant
852 mitochondria. Mitochondrion. 2017;33:22-37.
853 41. Morgenstern M, Stiller SB, Lübbert P, Peikert CD, Dannenmaier S, Drepper F, et al.
854 Definition of a High-Confidence Mitochondrial Proteome at Quantitative Scale. Cell reports.
855 2017;19(13):2836-52.
856 42. Putignani L, Tait A, Smith HV, Horner D, Tovar J, Tetley L, et al. Characterization of a
857 mitochondrion-like organelle in Cryptosporidium parvum. Parasitology. 2004;129(Pt 1):1-18.
858 43. Pryszcz LP, Huerta-Cepas J, Gabaldón T. MetaPhOrs: orthology and paralogy
859 predictions from multiple phylogenetic evidence using a consistency-based confidence
860 score. Nucleic acids research. 2011;39(5):e32.
861 44. Bender A, van Dooren GG, Ralph SA, McFadden GI, Schneider G. Properties and
862 prediction of mitochondrial transit peptides from Plasmodium falciparum. Molecular and
863 biochemical parasitology. 2003;132(2):59-66.

45. Alberts BJ, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. Molecular Biology of the Cell New York: Garland Science; 2002 [4th edition:[The Mitochondrion]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26894/.
46. Kozlowski LP. Proteome-pI: proteome isoelectric point database. Nucleic acids research. 2017;45(D1):D1112-d6.
47. Patrickios CS. Polypeptide Amino Acid Composition and Isoelectric Point: 1. A Closed-Form Approximation. Journal of Colloid and Interface Science. 1995;175(1):256-60.
48. Otto TD, Böhme U, Jackson AP, Hunt M, Franke-Fayard B, Hoeijmakers WA, et al. A comprehensive evaluation of rodent malaria parasite genomes and gene expression. BMC biology. 2014;12:86.
49. Matz JM, Kooij TW. Towards genome-wide experimental genetics in the in vivo malaria model parasite Plasmodium berghei. Pathogens and global health. 2015;109(2):46-60.
50. Deligianni E, Morgan RN, Bertuccini L, Kooij TW, Laforge A, Nahar C, et al. Critical role for a stage-specific actin in male exflagellation of the malaria parasite. Cellular microbiology. 2011;13(11):1714-30.
51. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological-image analysis. Nature methods. 2012;9(7):676-82.
52. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015.
53. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, et al. gplots: Various R Programming Tools for Plotting Data. 2016.
54. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016.
55. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics (Oxford, England). 2005;21(20):3940-1.
56. Wickham H. scales: Scale Functions for Visualization. 2017.
57. Wickham H. Reshaping data with the reshape package. Journal of Statistical Software. 2007;21(12).
58. Takeda K, Komuro Y, Hayakawa T, Oguchi H, Ishida Y, Murakami S, et al. Mitochondrial phosphoglycerate mutase 5 uses alternate catalytic activity as a protein serine/threonine phosphatase to activate ASK1. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(30):12301-5.
59. Lo SC, Hannink M. PGAM5 tethers a ternary complex containing Keap1 and Nrf2 to mitochondria. Experimental cell research. 2008;314(8):1789-803.
60. Ma K, Chen G, Li W, Kepp O, Zhu Y, Chen Q. Mitophagy, Mitochondrial Homeostasis, and Cell Fate. Frontiers in cell and developmental biology. 2020;8:467.
61. Lo SC, Hannink M. PGAM5, a Bcl-XL-interacting protein, is a novel substrate for the redox-regulated Keap1-dependent ubiquitin ligase complex. The Journal of biological chemistry. 2006;281(49):37893-903.
62. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology. 2001;305(3):567-80.
63. 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.

910 64. Ude S, Lassak J, Starosta AL, Kraxenberger T, Wilson DN, Jung K. Translation
911 elongation factor EF-P alleviates ribosome stalling at polyproline stretches. Science (New
912 York, NY). 2013;339(6115):82-5.
913 65. Yanagisawa T, Sumida T, Ishii R, Takemoto C, Yokoyama S. A paralog of lysyl-tRNA
914 synthetase aminoacylates a conserved lysine residue in translation elongation factor P.
915 Nature structural & molecular biology. 2010;17(9):1136-43.
916 66. Fu X, Wang Y, Shao H, Ma J, Song X, Zhang M, et al. DegP functions as a critical
917 protease for bacterial acid resistance. The FEBS journal. 2018;285(18):3525-38.
918 67. Krojer T, Garrido-Franco M, Huber R, Ehrmann M, Clausen T. Crystal structure of
919 DegP (HtrA) reveals a new protease-chaperone machine. Nature. 2002;416(6879):455-9.
920 68. McDonough MA, Kavanagh KL, Butler D, Searls T, Oppermann U, Schofield CJ.
921 Structure of human phytanoyl-CoA 2-hydroxylase identifies molecular mechanisms of
922 Refsum disease. The Journal of biological chemistry. 2005;280(49):41101-10.
923 69. Xu Q, Rawlings ND, Chiu HJ, Jaroszewski L, Klock HE, Knuth MW, et al. Structural
924 analysis of papain-like NlpC/P60 superfamily enzymes with a circularly permuted topology
925 reveals potential lipid binding sites. PloS one. 2011;6(7):e22013.
926 70. Mary C, Scherrer A, Huck L, Lakkaraju AK, Thomas Y, Johnson AE, et al. Residues in
927 SRP9/14 essential for elongation arrest activity of the signal recognition particle define a
928 positively charged functional domain on one side of the protein. RNA (New York, NY).
929 2010;16(5):969-79.
930 71. Janda CY, Li J, Oubridge C, Hernández H, Robinson CV, Nagai K. Recognition of a
931 signal peptide by the signal recognition particle. Nature. 2010;465(7297):507-10.
932 72. Hilario E, Li Y, Niks D, Fan L. The structure of a Xanthomonas general stress protein
933 involved in citrus canker reveals its flavin-binding property. Acta crystallographica Section D,
934 Biological crystallography. 2012;68(Pt 7):846-53.
935 73. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018
936 update. Nucleic acids research. 2018;46(W1):W200-w4.
937 74. Posayapisit N, Songsungthong W, Koonyosying P, Falade MO, Uthaipibull C,
938 Yuthavong Y, et al. Cytochrome c and c1 heme lyases are essential in Plasmodium berghei.
939 Molecular and biochemical parasitology. 2016;210(1-2):32-6.
940 75. Babbitt SE, San Francisco B, Bretsnyder EC, Kranz RG. Conserved residues of the
941 human mitochondrial holocytochrome c synthase mediate interactions with heme.
942 Biochemistry. 2014;53(32):5261-71.
943 76. Wang KT, Desmolaize B, Nan J, Zhang XW, Li LF, Douthwaite S, et al. Structure of the
944 bifunctional methyltransferase YcbY (RlmKL) that adds the m7G2069 and m2G2445
945 modifications in Escherichia coli 23S rRNA. Nucleic acids research. 2012;40(11):5138-48.
946 77. Feagin JE, Harrell MI, Lee JC, Coe KJ, Sands BH, Cannone JJ, et al. The fragmented
947 mitochondrial ribosomal RNAs of Plasmodium falciparum. PloS one. 2012;7(6):e38320.
948 78. van Dam TJP, Kennedy J, van der Lee R, de Vrieze E, Wunderlich KA, Rix S, et al.
949 CiliaCarta: An integrated and validated compendium of ciliary genes. PloS one.
950 2019;14(5):e0216705.
951 79. Smith DG, Gawryluk RM, Spencer DF, Pearlman RE, Siu KW, Gray MW. Exploring the
952 mitochondrial proteome of the ciliate protozoon Tetrahymena thermophila: direct analysis
953 by tandem mass spectrometry. Journal of molecular biology. 2007;374(3):837-63.
954