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], Joeri van Strien[1], 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 mitochondrial 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 445 proteins with a sensitivity of 87% and a specificity of 97%. 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.

42

43 **Introduction**

44 Although all mitochondria evolved from a single endosymbiotic event (1), they display a large
45 variety among the eukaryotes. For example, while in most species the mitochondrial
46 organelle still retains its genome, the number of proteins encoded varies from 100 in
47 *Andalucia godoyi* (2) to three in myzozoan species like *Plasmodium falciparum* that only
48 encode cytochrome b, cytochrome c oxidase subunits 1 and 3, and two fragmented rRNAs
49 (3). Moreover, mitochondria-like organelles of a variety of anaerobic species have lost the
50 organellar genome altogether (4). Also mitochondrial proteomes vary broadly in size, ranging
51 from e.g. ~1500 different proteins in mammalian mitochondria (5) to a few members of only
52 a single biochemical pathway found in some mitosomes (6). Within this variety, the
53 *P. falciparum* mitochondrion, based on the size of its genome, the size of the nuclear genome,
54 and the complexity of its oxidative phosphorylation, is predicted to have a relatively small
55 proteome. Although several mitochondrial pathways have already been identified in
56 *P. falciparum* (7), and databases like PlasmoDB contain valuable information about
57 mitochondrial localization of individual proteins (8), a complete resource that weighs and
58 combines all the relevant information about possible mitochondrial localization is lacking.

59 *Plasmodium* mitochondria are particularly interesting because their function is
60 variable between different life-cycle stages. In the sexual blood stage, functional oxidative
61 phosphorylation is essential for colonization and development inside the mosquito hosts (9,
62 10), while in asexual blood stages the only essential function of the respiratory chain is to
63 recycle ubiquinone for pyrimidine biosynthesis (11). This is further highlighted by the
64 observation that in asexual blood-stage parasites no cristae are detectable in the
65 mitochondrial membrane, while in gametocyte mitochondria cristae-like structures that
66 typically accumulate respiratory chain complexes are observed (10, 12). A recent
67 complexome profiling study demonstrated that the level of complex V components in asexual
68 blood-stage parasites was only 3% of the level in gametocytes (12). Another intriguing aspect
69 of mitochondria in *Plasmodium* and closely related species is that they have a rather unique
70 FeS cluster assembly pathway (13). From this pathway the frataxin protein, an FeS assembly
71 protein that occurs even in mitosomes (14) appears to be missing in *Plasmodium* (15).

72 Knowledge on the mitochondrial proteome could provide viable targets for the
73 development of new drugs against this deadly parasite, as apicomplexan mitochondrial

74 proteomes contain unique proteins that are not present in the animal hosts (16). Recently for
75 example, an unusual prohibitin-like protein, PHBL, was identified in *Plasmodium berghei* that
76 might provide a transmission-blocking drug target as it is unique to Myzozoa and *phbl*$^{-}$
77 parasites fail to transmit (17). Furthermore, respiratory protein complex III has proven to be
78 a good drug target (18) and inhibitors to the mitochondrial enzymes dihydroorotate
79 dehydrogenase (DHODH) and NADH type II oxidoreductase (PfNDH2) have been identified
80 (19).

81 For years, it has been difficult to perform proteomics experiments to identify the
82 *Plasmodium* mitochondrial proteome due to the lack of organelle isolation methods that
83 reliably separate mitochondria from the apicoplast (20). The definition of mitochondrial
84 functions relied heavily upon homology studies (7, 21) and microscopic examination of
85 individual proteins (Supplemental Table 1). With the use of two biotin tagging approaches,
86 422 mitochondrial matrix proteins were identified in *Toxoplasma gondii*, a species related to
87 *Plasmodium* that also has an apicoplast and a mitochondrion (16). Besides these
88 mitochondria-targeted approaches, hyperLOPIT, a whole-cell biochemical fractionation
89 technique, defined the subcellular localization of the *T. gondii* proteome. This provided
90 among others a set of 220 soluble and 168 membrane-bound *T. gondii* mitochondrial proteins
91 (22). Alongside these proteomic data, there is a wealth of other data sources providing
92 information on mitochondrial localization in *P. falciparum*. The challenge remains to use and
93 combine these heterogeneous datasets to reliably predict the mitochondrial proteome.

94 Here, we integrated eight datasets containing proteomic, gene expression, orthology
95 to mitochondrial proteins from other species, phylogenetic distribution, and amino acid
96 sequence data in a Bayesian manner as has been applied before such as for the definition of
97 the mitochondrial proteome of humans (13) and of developmental stage specific proteins in
98 Plasmodium(23, 24). We used a naïve Bayesian classifier to combine the data. This method
99 exploits the relative strengths of the various datasets while maintaining transparency about
100 the contribution of each dataset to the final Bayesian posterior probability score for being
101 mitochondrial. For each protein, the contribution of each dataset to the prediction was
102 determined using gold standards of *Plasmodium* proteins that are either known to be
103 mitochondrial or non-mitochondrial. After data integration, we obtained a list of 445 P.
104 falciparum mitochondrial proteins. We experimentally validated these predictions with seven

105 proteins of varying probability scores, and collectively this predicted proteome indicates that
106 the *P. falciparum* mitochondria are unique relative to mitochondria of model organisms.
107
108 **Materials and Methods**
109
110 **Animal experiments.** All animal experiments were performed in accordance with the Dutch
111 Experiments on Animals Act (Wod), the Directive 2010/63/EU from the European Union and
112 the European ETS 123 convention and approved by the Radboud University Animal Welfare
113 Body (IvD) and Animal Experiment Committee (RUDEC; 2015-0142), and the Central Authority
114 for Scientific Procedures on Animals (CCD; AVD103002016424). In this study, we used
115 outbred male and female NMRI mice (Envigo).
116
117 **Nuclear encoded reference proteome.** All datasets used in this study were mapped to the
118 *P. falciparum* 3D7 reference proteome version 3.1 from the Sanger Institute, downloaded
119 December 2017. This version, including isoforms, contains 5431 proteins encoded by 5357
120 genes. As we are interested in the proteins encoded by nuclear genes, all apicoplast and
121 mitochondrial encoded proteins were excluded, leaving a reference of 5324 genes. When a
122 dataset contains informational data points for two or more protein isoforms of one gene, the
123 value chosen for that gene was the one that comes closest to the expected values for
124 mitochondrial proteins in that specific dataset.
125
126 **Assembly of positive and negative lists for benchmarking.** Bayesian data integration depends
127 on gold standards of proteins known to be mitochondrial or non-mitochondrial. We
128 constructed two gold standards to assess the predictive value of the individual datasets.
129 Furthermore, we built two alternative sets for the construction of the CLIME (co-evolution)
130 and WICCA (co-expression) datasets to prevent that the standards chosen to train those two
131 predictors were also used to evaluate them. Thus we avoid circular arguments in the data
132 integration. These four sets are unique, have no overlapping genes, and are available in
133 Supplemental Table 1. To construct these lists, we commenced with a systematic review of
134 all literature available via PubMed (as per 30-08-2018) using the following broad search string
135 "(mitochondri* OR apicoplast OR plastid) AND (plasmodium OR malaria)". Given the intimate
136 functional and physical relation between mitochondrion and apicoplast, we reasoned that

137 the positive gold standard list should only include genes encoding proteins that had been
138 unambiguously and singularly assigned to the mitochondrion through fluorescence
139 microscopy and co-localization with confirmed mitochondrial markers or via immune-
140 electron microscopy (54 genes). Any proteins for which multiple studies showed different
141 results were excluded. To facilitate discrimination from the apicoplast, we also made a gold
142 standard apicoplast list (72 genes). Finally, all proteins that were unambiguously non-
143 mitochondrial as demonstrated in the same papers (170 genes) were combined with the
144 apicoplast-positive list to form the basis for the negative gold standard. This negative set was
145 then complemented with an additional 346 non-overlapping genes encoding non-apicoplast
146 and non-mitochondrial proteins identified through extensive literature review and published
147 online by the Ralph lab (25). The 588 genes encoding non-mitochondrial proteins were
148 randomly assigned to the negative gold standard (294 genes) that was used for the Bayesian
149 data integration and the alternative negative gold standard (294 genes) that was used to train
150 the CLIME and WICCA approaches. For the latter purpose, an alternative positive standard
151 was compiled of 146 genes, non-overlapping with the first gold standard, but associated with
152 one or more mitochondrial GO-terms (GO:0000275, GO:0005739-43, GO:0005746,
153 GO:0005750, GO:0005753, GO:0005758, GO:0005759, GO:0006122, GO:0006839,
154 GO:0006850, GO:0031966, GO:0033108, GO:0042775, GO:0044429, GO:0044455).
155
156 **Datasets to predict mitochondrial and non-mitochondrial proteins.** To predict mitochondrial
157 proteins and non-mitochondrial proteins, we collected and constructed eight datasets (Table
158 1) that contain features typical of either mitochondrial or apicoplast proteins. Some proteins,
159 like lipoate protein ligase 2 (PF3D7_0923600) (26) and serine hydroxymethyltransferase
160 isoforms (PF3D7_1235600) (27), are dual localized to the mitochondrion and apicoplast,
161 highlighting the need for the second category of datasets. Each dataset is described in detail
162 below, including how the dataset was processed to obtain per *P. falciparum* protein a single
163 score relevant to whether the protein is mitochondrial or not.
164
165
166
167
168
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 'omics data					
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
Non- Apicoplast proteins (28)	BioID, gene expression	5324	93.6	94.4	5.4E-01
Datasets based on ortholog evidence					
hyperLOPIT (TAGM.MCMC.joint.mitochondrion_max ≥ 4.78E-05) (22)	hyperLOPIT	2246	15	95.2	9.3E-32
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
Isoelectric point (Patrickios pI ≥ 4.2)*(28)*	pI	5324	55.1	83.3	1.1E-05

172

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

176 overrepresentation of identified gold standard genes compared to random
177 expectation.Notice that in the "non-Apicoplast proteins" and in the "Absence of a
178 Cryptosporidium ortholog" these P values reflect the over representation of mitochondrial
179 proteins among the large sets of proteins to which these features apply.
180
181 *Dataset based on 'omics data*
182 **Co-expression.** Co-expressed genes tend to code for proteins that functionally interact. For
183 mammals, co-expression analyses have been successfully used in the discovery of
184 mitochondrial proteins (5, 31). To determine if the expression pattern of a *Plasmodium* gene
185 correlates with mitochondrial protein coding genes, we developed a co-expression tool:
186 WeIghted Co-expression Calculation tool for plAsmodium genes (WICCA, available at
187 https://wicca.cmbi.umcn.nl/). In short, the tool uses the co-expression of a group of input
188 genes with each other to weigh 83 microarray experiments for their predictive value for that
189 group. Datasets that show high co-expression of the input genes are more likely to be relevant
190 for the system that the input genes are part of, and receive a higher weight. It then uses these
191 weights to combine the expression datasets and calculate one co-expression score per gene
192 with the input system, in this case, one score per gene for the level of co-expression with
193 mitochondrial protein coding genes. We used the co-expression ranking obtained with the
194 alternative positive standard of 146 genes as input for WICCA to weigh the co-expression
195 datasets. The methods of this tool were based on the WeGET method
196 (http://weget.cmbi.umcn.nl/, (32)), methodological details on WICCA can be found in the
197 supplement.
198
199 **Bio-ID.** In Bayesian data integration, "negative data sets" that e.g. predict that a protein is
200 non-mitochondrial can be as valuable as positive ones**.** We used the set of 346 apicoplast
201 proteins that are based on BioID data and gene expression data analyzed with a Neural
202 Network (28) to reduce apicoplast protein contamination among the top scoring proteins.
203
204 *Datasets based on ortholog evidence*
205 **hyperLOPIT.** Hyperplexed Localization of Organelle Proteins by Isotope Tagging (hyperLOPIT)
206 is a proteomics technique that analyzes protein distributions upon biochemical fractionation.
207 It enables the identification of the subcellular localization of thousands of proteins (33, 34).

208 Barylyuk *et al.* used this technique on the apicomplexan *T. gondii* and, among others,
209 classified proteins identified in all three hyperLOPIT experiments to be part of the
210 mitochondrial soluble matrix and mitochondrial membrane. This classification was based on
211 t-augmented Gaussian mixture models (TAGM) in combination with maximum a posteriori
212 prediction (TAGM-MAP) and Markov-chain Monte-Carlo (TAGM-MCMC) methods (22).
213 Marker proteins used for the classification were arbitrarily set to 1 in the class that they
214 belong to and to 0 for all other classes. Using BLAST (35), or if that produced no homologs
215 HHpred (36), in combination with selecting for best bidirectional hits, we determined a list of
216 orthologs between *T. gondii* and *P. falciparum*. The 3,832 *T. gondii* proteins that got a TAGM-
217 MCMC probability for mitochondrial soluble matrix and mitochondrial membrane were
218 mapped to the corresponding *P. falciparum* ortholog when available. This resulted in a list of
219 2246 *P. falciparum* identifiers with two TAGM-MCMC probabilities, one for mitochondrial
220 matrix localization and one for mitochondrial membrane localization. As the input datasets
221 for the Bayesian integration should be independent, both TAGM-MCMC probabilities cannot
222 be included as separate datasets. Therefore the maximum TAGM-MCMC probability (so
223 either the soluble or the membrane value) for each protein was used to create one input
224 hyperLOPIT dataset.

225 **Mitochondrial ortholog.** We also used orthologs from more distantly related species
226 than *T. gondii*, as also at larger phylogenetic distances subcellular localizations tend to be
227 conserved (37). Using BLAST (35) (all versus all blast for proteomes with e-value of 100) in
228 combination with OrthoMCL (38) (percent match cut of 50, e-value exponent cut off -5) to
229 detect orthologs, or if that produced no result, HHpred (36) (default settings, e-value cut off
230 0.01, three iterations, only best-bi-directional hits included), one-to-one orthologs between
231 *P. falciparum* and either *Homo sapiens*, *Arabidopsis thaliana* or *Saccharomyces cerevisiae*
232 were determined. When the orthologous protein of at least one species was annotated to be
233 mitochondrial in a published mitochondrial compendium (*H. sapiens* MitoCarta2.0 (5), *A.*
234 *thaliana* (39), *S. cerevisiae* (40)), then that *P. falciparum* protein was included in the
235 mitochondrial ortholog dataset.

236 **Evolutionary inference.** Clustered by inferred models of evolution (CLIME) uses
237 homology across model species (138 eukaryotic species and 1 prokaryotic outgroup) to
238 identify evolutionary conserved clusters. We downloaded the pre-computed CLIME analyses
239 of *P. falciparum* genes from http://gene-clime.org/ (29) and mapped the IDs to our reference

240 proteome. We used the CLIME matrix listing the presence/absence of orthologs of
241 *P. falciparum* in other species, with the alternative positive and negative sets as training data,
242 to train a perceptron with two hidden layers to obtain a single score per protein (details in
243 supplement). The resulting output scores on our test set formed the CLIME input for the
244 Bayesian integration.

245 **Absence of a *Cryptosporidium* ortholog.** Like *Plasmodium*, the genus *Cryptosporidium*
246 belongs to the phylum of Apicomplexa. *Cryptosporidium* species lack an apicoplast and
247 contain only a remnant mitochondrial-like organelle (41). Therefore, it can be expected that
248 apicoplast proteins and many mitochondrial proteins will not have an ortholog in
249 *Cryptosporidium*. MetaPhOrs is an online tool that contains orthology and paralogy
250 predictions obtained from multiple phylogenetic trees (42) and was used to assess orthology
251 for *P. falciparum* genes in three *Cryptosporidium* species (*hominis, parvum*, and *muris*). We
252 calculated a combined score by multiplying two MetaPhOrs metrics, one for the confidence
253 level (see below) and one for the number of hits between the four species, and used this
254 combined score as input for the Bayesian integration. In detail, the MetaPhOrs results include
255 a consistency score (ranging from 0 with no overlap to 1 if all trees contain the same protein
256 relationship/orthology information) and an evidence level for the number of independent
257 databases (the theoretical maximum is 13, but for our species combination it was four). The
258 multiplication of consistency and evidence level results in an arbitrary score (range 0-4) that
259 indicates the confidence in calling two proteins orthologs. *P. falciparum* proteins with a high
260 score are more likely to have an ortholog in *Cryptosporidium*. Notice that there is some
261 overlap with the Evolutionary inference as that includes one *Cryptosporidium* species.
262 Nevertheless, *Cryptosporidium* species show quite some variation in the complexity of their
263 mitosomes and the overlap in the predictions is limited (see below).
264

265 *Datasets based on amino acid sequence evidence*

266 **Mitochondrial targeting signal.** The canonical mitochondrial import system requires an
267 amphipathic α-helical N-terminal targeting sequence. As *Plasmodium* has a distinct amino
268 acid usage pattern (43), a *P. falciparum* specific tool, PFMpred (30), was chosen to predict
269 mitochondrial localization based on the amino acid sequence. This is a support vector
270 machine (SVM) based tool and was used in split-amino-acid-composition mode (Matthews
271 correlation coefficient 0.73) that allows separate calculations for the N-terminus, C-terminus,

272 and remainder of the protein. The tool reports one SVM-score per protein and indicates
273 whether the protein is predicted to be mitochondrial or not. The SVM-scores per gene are
274 used as the mitochondrial targeting signal dataset.

275 **Isoelectric point.** Nuclear encoded mitochondrial proteins need to cross the
276 negatively charged mitochondrial membrane and need to function properly inside the
277 mitochondrial environment that has a slightly higher pH compared to the cytosol (44). A
278 density plot (Supplemental Fig 1) by Patrikios isoelectric points (45) of the human
279 mitochondrial proteome (MitoCarta2.0 (5)) showed that mitochondrial proteins on average
280 have a higher pI than the remainder of the proteome. Isoelectric points for the *P. falciparum*
281 proteome were calculated using the Patrickios algorithm (46) and directly used as pI scores.
282

283 **Bayesian integration to predict mitochondrial proteins.** Using the eight datasets described
284 above and the gold standard evaluation sets, a mitochondrial score for each *P. falciparum*
285 reference protein was calculated. This score is the logarithm of the odds that a protein is
286 localized in the mitochondrion relative to the protein being localized somewhere else in the
287 cell. For categorical data (mitochondrial ortholog, apicoplast targeting signal), the dataset is
288 separated into bins that represent the two categories. Continuous data (the other six
289 datasets) were binned in a systematic way using a custom python script
290 (https://github.com/JordyCoolen/Binning) (settings: --bins 0 --score 1). In short, this script
291 optimizes for each dataset the distribution of the bins such that the sum of log ratio scores of
292 all bins combined (see below) is close to zero, as this will result in the best separation of
293 positive and negative bins. The number of bins was varied to a maximum of five bins to
294 achieve the most optimal separation without expanding the number such that there are too
295 few proteins per bin to obtain reliable estimates of the log odds of that bin.

296 The fractions of gold standard proteins per bin are determined. The log ratio of these
297 fractions determines the score for all other proteins in that respective bin. The mitochondrial
298 score is based on the sum of log ratios of the individual datasets and is calculated as follows:
299

300 $$mitochondrial\ score$$

301 $$= log_2\left(\frac{P_{mitochondrial}}{P_{non-mitochondrial}}\right) + \sum_{i=1}^{n} log_2\left(\frac{P(data_i|mitochondrial)}{P(data_i|non-mitochondrial)}\right)$$

302

303 $$with\ \frac{P(data_i|mitochondrial)}{P(data_i|non-mitochondrial)} = \frac{mitochondrial_pos_i}{mitochondrial_neg_i}$$

304

305 where mitochondrial_pos and mitochondrial_neg are the fractions of the positive and
306 negative gold standard genes in sample i respectively. If there were no gold standard genes
307 found in a certain bin, that positive or negative gold standard fraction was set to 0.5/'total
308 number of negative set genes in the complete dataset' to prevent division by zero and allow
309 calculation of the log ratio. The O_{prior}, $\log_2(P_{mitochondrial}/P_{non\text{-}mitochondrial})$, is based on the
310 estimation that 536 proteins out of the 5357 protein coding genes (~10%) encode a
311 mitochondrial protein. This is a conservative estimate when compared to single cell species
312 like *Saccharomyces cerevisiae* with 16% mitochondrial proteins (47), or the more closely
313 related *T. gondii* where 15% of proteins that could confidently be mapped are mitochondrial
314 (48), while compared to all proteins that were mapped in that dataset the percentage of
315 mitochondrial proteins is 10%. Note that the O_{prior} only affects the score per protein, it does
316 not affect the relative ranking of potential mitochondrial proteins.

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

322

323 $$cFDR = \frac{1 - specificity}{1 - specificity + sensitivity \cdot O_{prior}}$$

324

325

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

333

334 **Intraspecies homologs.** The human mitochondrial proteome has expanded since the
335 divergence from yeast mainly due to gene duplications that create mitochondrial paralogs
336 (37). It is therefore interesting to see if the predicted *P. falciparum* mitochondrial proteins
337 have intraspecies homologs, and whether we can identify mitochondrial gene families. With
338 HHpred (36), using default settings and a 1E-5 E-value cut-off, the homologs within the
339 genome of *P. falciparum* were determined and included as a column in Supplemental Table
340 2.

341

342 **Validation of candidate proteins.** To validate the predictions, we selected 14 genes with
343 unusual or novel mitochondrial functions within the top 295 genes that fell within the 25%
344 cFDR cut-off in the initial data integration. Note that during the project, the number of
345 predicted mitochondrial proteins increased to 445 by including better orthology prediction
346 and a better data set of likely apicoplast proteins. Of these 14 genes, 13 were selected based
347 on maximum transcript levels during asexual blood-stage development of *P. berghei* strain
348 ANKA cl15cy1 that exceeded 100 FPKM (49). In addition, we selected one particularly
349 interesting gene, PF3D7_1004900 that had a lower transcript level. We used experimental
350 genetics (50) to generate *P. berghei* lines expressing 3xHA-tagged copies of the selected
351 proteins and colocalized them with an established mitochondrial marker (17) to determine
352 their subcellular localization by immunofluorescence microscopy. Details on plasmid
353 construction (successful for 13 out of 14 targets, excluding PF3D7_1142800) and the
354 generation of the lines can be found as supplemental information. To assess colocalization,
355 the freshly harvested transgenic parasites were allowed to settle on a poly-L-lysine coated
356 coverslip for 10 minutes and then fixed with 4% EM-grade paraformaldehyde (Fisher
357 Scientific) and 0.0075% EM-grade glutaraldehyde (Sigma-Aldrich) in microtubule stabilizing
358 buffer (MTSB, 10 mM MES, 150 mM NaCl, 5 mM EGTA, 5 mM glucose, 5 mM MgCl2 pH 6.9)
359 for 20 minutes (51). Next, parasites were permeabilized with 0.1% Triton X-100 in PBS for 10
360 minutes. Samples in which we imaged mOrange and mCherry were hereafter stained with
361 DAPI (1:300) for 15 minutes and mounted on a microscope slide with VectaShield (Vector
362 laboratories). All other samples were blocked in 3% FCS-PBS for 1 hour. Samples were
363 incubated overnight at 4°C with rat anti-HA (1:500, ROAHAHA Roche) and chicken anti-GFP
364 (1:1000, Thermo Fisher) antibodies and for 1 hour with 1:500 goat anti-rat Alexa fluor™ 594

365 (Invitrogen) and goat anti-chicken Alexa fluor™ 488 (Thermo Fisher) conjugated antibodies at
366 room temperature. Nuclei were stained with DAPI (1:300) for 15 minutes and coverslips were
367 mounted on a microscope slide with VectaShield (Vector laboratories). Samples were imaged
368 with an SP8 confocal microscope (Leica, 63x oil lens) or LSM900 confocal microscope (Zeiss,
369 63x oil lens). Images were processed minimally and similarly with Fiji (52).
370
371 **Tools for data analysis.** Plots, statistics, and calculations were performed with the R statistical
372 package (53) and additional packages gplots (54), ggplot2 (55), ROCR (56), scales (57) and
373 reshape (58). The separation of the datasets into bins was performed using a python script
374 (see above).
375
376 **Results**
377 We integrated eight features of proteomic, gene expression, orthology and amino acid
378 composition data, to predict mitochondrial localization in *P. falciparum*. All the individual
379 features had a predictive value for mitochondrial localization (Fig 1A and 1B), including, as
380 expected, a negative score for the "*Cryptosporidium* ortholog" and the "apicoplast" datasets
381 (Supplemental Fig 2A). Notice that for the co-expression feature it is the bin with the low-rank
382 values that contains most mitochondrial proteins because this bin contains the proteins with
383 the highest co-expression values. The best performing input datasets are the ones that
384 translate mitochondrial localization between species: the hyperLOPIT dataset and the
385 mitochondrial ortholog dataset, followed by the mitochondrial targeting signal and co-
386 expression datasets (Fig 1B). Correlation analyses showed that the features are largely
387 independent off each other, allowing us to use naïve Bayesian data integration (Supplemental
388 Fig 2B). The highest correlations were observed between features that examine the
389 phylogenetic distribution of the proteins: CLIME and *Cryptosporidium* ortholog datasets. By
390 integrating all data, we achieved both a better sensitivity and specificity in predicting
391 mitochondrial proteins with an AUC of 0.959 than we achieved for individual data sets (Fig
392 1B). The performance of the hyperLOPIT dataset on its own approximates the integrated
393 prediction with an AUC of 0.953 (Fig 1B). However, this dataset contains information on only
394 2246 (42%) of the nuclear encoded *Plasmodium* proteins, highlighting the value of additional
395 datasets and their integration to predict a complete mitochondrial proteome. The distribution
396 of the log-odds ratios (Fig 1C) showed that the mitochondrial score separated the positive

397 and negative gold standards well from each other and that there are additional proteins with
398 a high score that were not part of either, which are potential new mitochondrial proteins. We
399 use a cFDR of 25% (Fig 2A) as a threshold for highly probable new mitochondrial proteins and
400 identify 445 proteins (Fig 2B) with a sensitivity of 87% and a specificity of 97%. The ranked list
401 of the complete nuclear encoded proteome is available in Supplemental Table 2.
402
403 *Comparison of the Toxoplasma and Plasmodium mitochondrial proteomes*.
404 Overall, the hyperLOPIT dataset contains 388 *T. gondii* mitochondrial proteins with a TAGM-
405 MCMC location probability of 99% or higher. Of those, 296 have a *P. falciparum* ortholog and
406 of these orthologs, we find 273 in our predicted mitochondrial proteome. Most proteins that
407 we predicted to be mitochondrial that were not supported by the *T. gondii* data are proteins
408 that do have orthologs in *T. gondii*, for which there were however no hyperLOPIT data (95
409 proteins). There are furthermore 38 proteins in the set of 445 that do not have *T. gondii*
410 orthologs. It should be noted that these do not include the six *P. falciparum* mitochondrial
411 ribosomal proteins that, based on Blast searches, were deemed to be absent from *T. gondii*
412 (59), as for those we could identify orthologs using HHpred (Table S2), underlining the
413 relevance of sensitive homology detection. Finally, there are 39 proteins of which orthologs
414 are present in the hyperLOPIT dataset but fell outside the 99% cut off for mitochondrial
415 localization that was used in that study, 26 of these did still have a mitochondrial localization
416 as the most probable location and 13 had a different predicted location in *T. gondii* (flagged
417 in Supplemental Table 2). Manual examination of those 13 inconsistencies reveals proteins
418 like PF3D7_1125300/mitochondrial RNA polymerase that in the hyperLOPIT data has been
419 observed in the nucleolus, but also PF3D7_1446400/pdhB that was predicted to be
420 mitochondrial based on orthology with pdhB in e.g. *H.sapiens* but is localized in the apicoplast
421 in *Plasmodium(60)*.
422

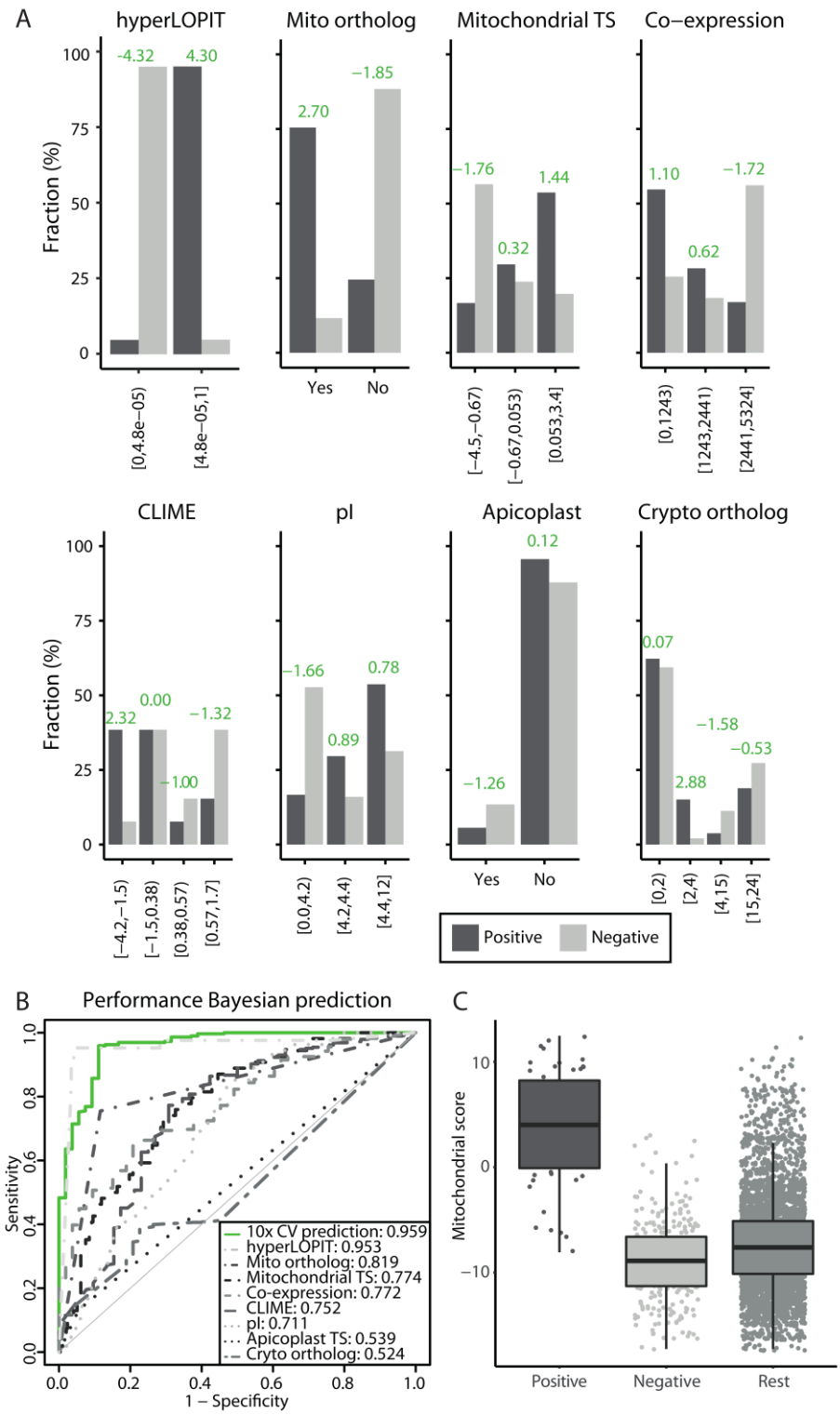
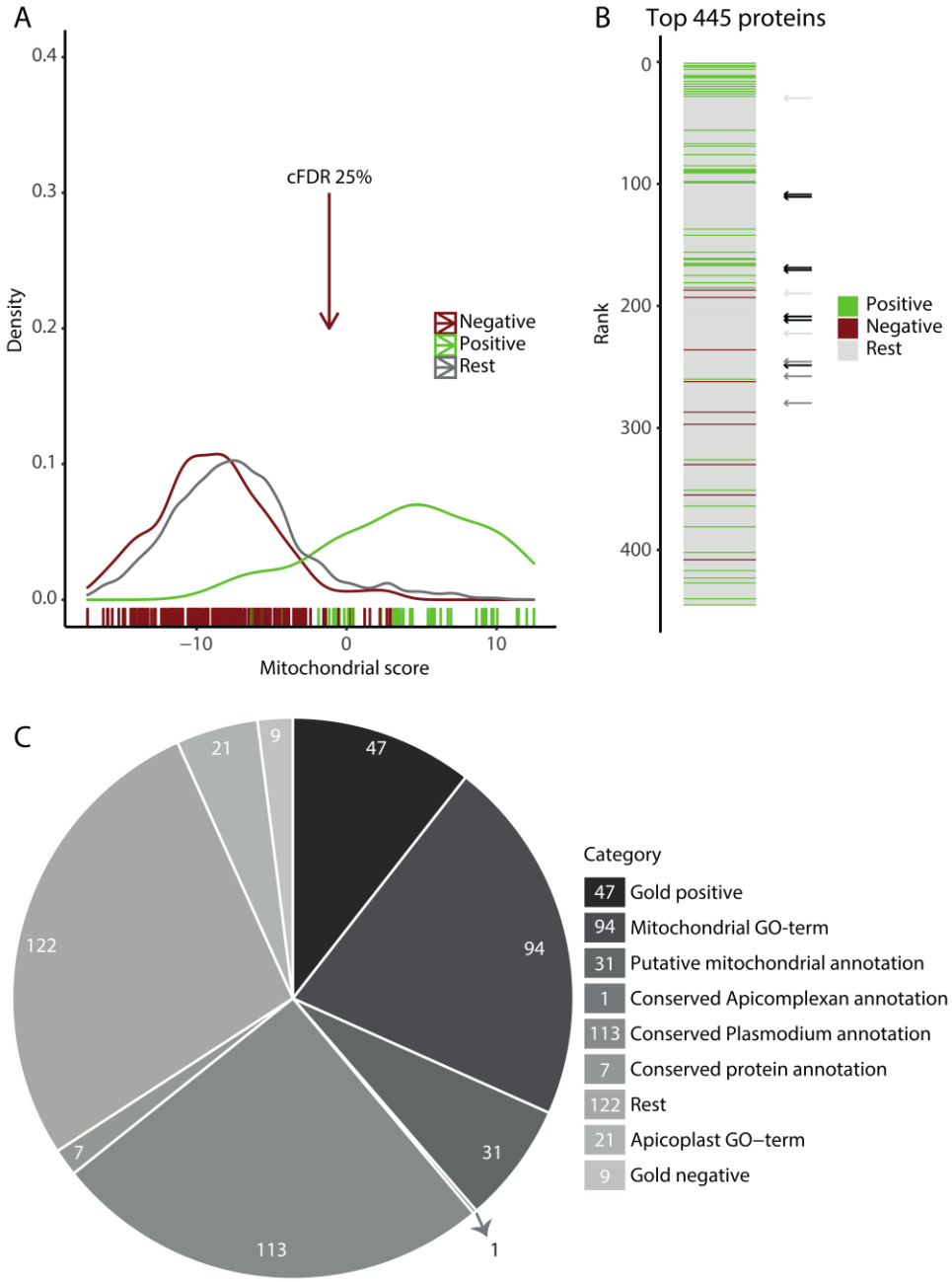


Fig 1. Predictive values of the individual datasets and the integrated ranked list. (A) Barplot of each dataset with the fraction of the gold standards on the y-axis and the bins on the x-axis. In green, the mitochondrial score per bin is indicated to show the predictive value of each bin. (B) ROC-curves comparing the performance to identify mitochondrial proteins of the prediction (10-fold cross-validated) to the performance of the individual datasets. For

429 comparison, the values for the area under each curve are indicated in the legend. (C) Boxplot
430 that visualizes the distribution of the gold-standard positives, negatives, and remainder of the
431 proteome (x-axis) over the calculated mitochondrial score of the prediction (y-axis).
432

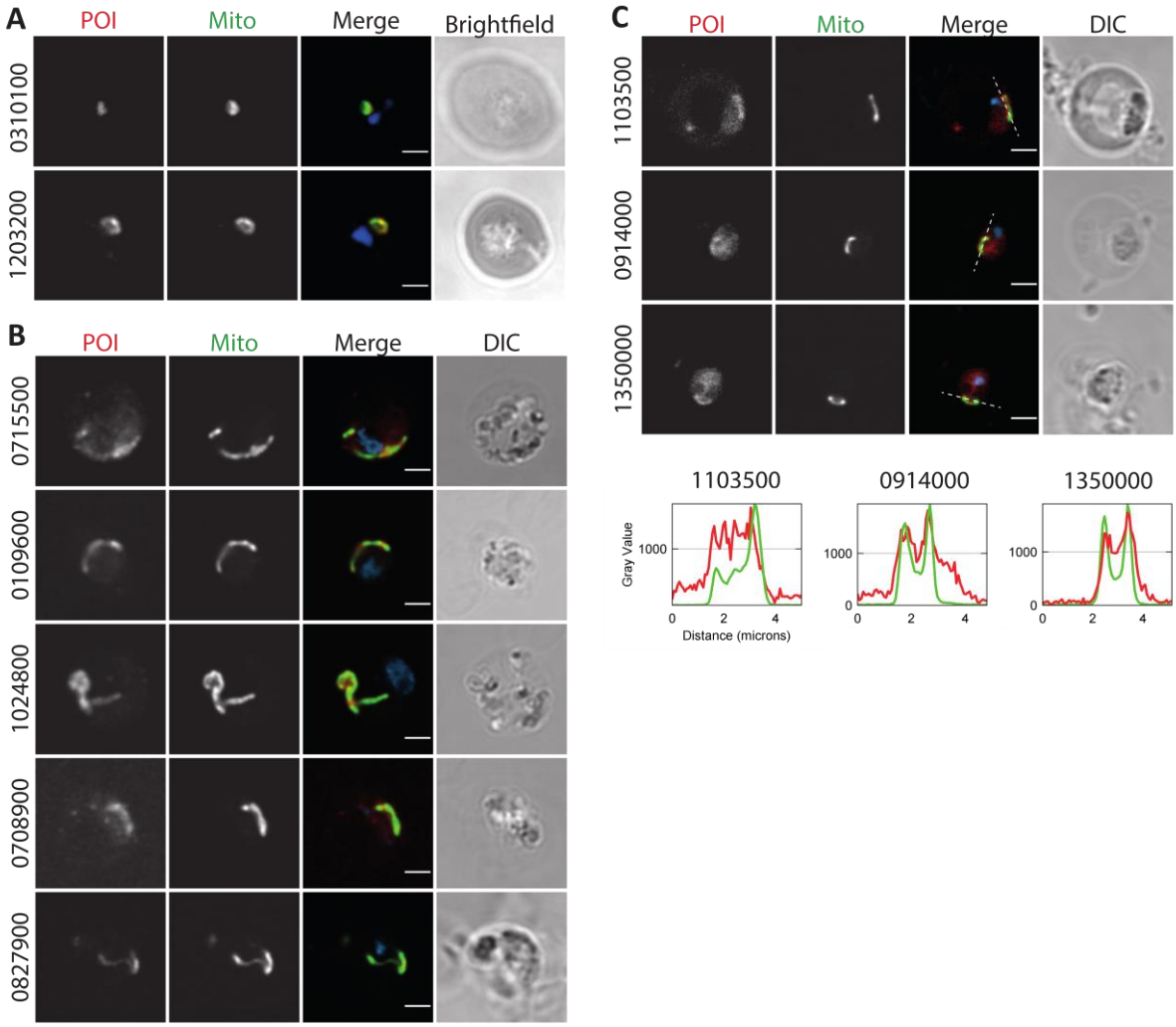


433
434

435 **Fig 2. Integration produces a ranked proteome with 445 likely mitochondrial proteins at a**
436 **25% cFDR-cut.** (A) Density plot of the mitochondrial score with a colored bar at the bottom
437 indicating the scores of individual gold standards. The arrow indicates the 25% cFDR cut-off.
438 (B) The top 445 proteins falling in the 25% cFDR cut-off ranked on their mitochondrial score

439 with color indicating if the protein is part of a gold standard. Black arrows indicated the
440 ranking of the seven candidate proteins that we experimentally confirmed to be
441 mitochondrial (Fig 3A and 3B), dark grey arrows indicate the ranking of the three proteins
442 with suggested mitochondrial localization (Fig 3C), and light grey arrows the ranking of the
443 three proteins with inconclusive results (Supplemental Fig 4B). (C) Categorical representation
444 of the 445 proteins identified at the 25% cFDR cut-off. First, genes were separated into
445 categories based on the gold standard, the remainder on mitochondrial and apicoplast GO
446 term, and subsequently the remainder on gene function annotations.
447
448 **Validation of candidates.** We aimed to validate the mitochondrial candidate list by tagging
449 orthologs of the selected unusual or unique predicted mitochondrial proteins in the efficient
450 transfection model *P. berghei* (see the section 'new mitochondrial proteins' for the *P.*
451 *falciparum* orthologs of these proteins). Mitochondrial localization was assessed by
452 fluorescent microscopy using an experimental genetic approach developed in our lab (17).
453 This method allows the endogenous tagging of a target protein with a combined fluorescent
454 and epitope tag, while simultaneously introducing a strong mitochondrial targeted GFP
455 marker by fusion of the promoter and N terminus of HSP70-3 (PBANKA_0914400) to GFP.
456 Initially, we tagged 6 proteins with a mOrange-3xHA tag. Live imaging revealed very poor and
457 undefined signals (data not shown). Next, we fixed blood samples and stained them with anti-
458 HA antibodies. Using this approach, only PBANKA_0310100 and PBANKA_1203200 were
459 convincingly shown to localize to the mitochondrion (Fig 3A), while other proteins
460 demonstrated mostly undetectable or undefined signals (Supplemental Fig 4A). Since many
461 of the selected proteins are relatively small compared to the mOrange-tag and are expected
462 to be imported into the mitochondrion, we anticipated that interference of the tag with
463 import or proper folding could lead to the observed weak and undefined signals. To assess
464 this, we selected two similarly small and abundantly expressed targets, PBANKA_0310100
465 (127 amino acids) and PBANKA_1024800 (144 amino acids), to test different tags. For
466 PBANKA_0310100, which was already successfully localized to the mitochondrion, these
467 included tags consisting of only the fluorescent protein mOrange or only the 3xHA epitope
468 tag, either with or without a linker sequence. For PBANKA_1024800, we also included the
469 combined mCherry-cMyc tag that has previously been used successfully to tag mitochondrial
470 proteins (17). We found that the fluorescent protein tag indeed interfered with the

471 localization of PBANKA_1024800, while the 3xHA tag, with or without a linker, supported a
472 mitochondrial localization (Supplemental Fig 5B). Surprisingly, using the mOrange-tag
473 without the subsequent HA-tag also led to an undefined rather punctuate mislocalization of
474 PBANKA_0310100 (Supplemental Fig 5B). Based on these observations, we decided to tag the
475 remaining 11 proteins with a linker-3xHA-tag (Fig 3B, 3C and Supplemental Fig 4B). Five of
476 these were also convincingly localized to the mitochondria (Fig 3B), while an additional three
477 proteins showed very weak staining patterns suggestive of a possible mitochondrial
478 localization (Fig 3C). The signals of the remaining three tagged proteins were unable to be
479 distinguished from the background and therefore no subcellular localization could be
480 assigned (Supplemental Fig 4B).



481

482 **Fig 3. Validation of mitochondrial candidate proteins.** Immunofluorescent analysis of tagged
483 candidate proteins of interest (POI) in *P. berghei* parasites (PBANKA_ID indicated vertically).
484 POI were stained with anti-HA antibody (red, first column), mitochondria were stained with
485 anti-GFP antibody binding to the organelle marker (Mito, green, second column). The DNA

486 was stained with DAPI (blue in merge). (A) Representative images of candidate proteins
487 tagged with a mOrange-3xHA tag, imaged on a SP8 confocal microscope (Leica). (B-C)
488 Representative images of candidate proteins tagged with a linker-3xHA tag, imaged on a
489 LSM900 confocal microscope (Zeiss). (C) Images with the brightest signal for three lowly
490 expressed candidate proteins. Pixel intensities of the POI and mitochondrial signal were
491 normalized and plotted over the line indicated in the merge panel. Scale bar, 2 µm.
492
493 **New mitochondrial proteins.** A categorical representation of the identified proteins (Fig 2C)
494 shows that 172 of the 445 proteins have a previous annotation of being mitochondrial as they
495 are either part of the gold standard, have a mitochondrial GO term, or are described as
496 putative mitochondrial protein. A smaller number of proteins, 30, were part of the negative
497 set or have an apicoplast GO-term & no mitochondrial GO term. The remaining 243 proteins,
498 of which 121 proteins are annotated as conserved hypothetical proteins, contain new
499 mitochondrial candidates.
500 For 27 of the 121 conserved proteins with an unknown function, we found a
501 mitochondrial ortholog using the sensitive homology detection tool HHpred (36) that can
502 improve their annotation (Supplemental Table 2). For example, the human and yeast
503 orthologs of PF3D7_0306000 (Bayesian rank 41) are both cytochrome b-c1 complex subunit
504 8, making it very likely that the *Plasmodium* protein is also cytochrome b-c1 complex subunit
505 8, as confirmed by recent complexome profiling (12). Other similar examples include
506 PF3D7_1357600 (rank 285) that is likely a mitochondrial member of the 39S ribosomal protein
507 family L53, and PF3D7_1309900 (rank 80) that is likely succinate dehydrogenase assembly
508 factor 2.
509 Besides the possibility to improve annotations using well-characterized mitochondrial
510 orthologs, the predicted 445 proteins contain biologically interesting proteins that can add
511 new mitochondrial functions to *P. falciparum* and related species. Therewith, the prediction
512 provides targets for experimental research to uncover new biology. These proteins can be
513 divided into five categories: (i) proteins with known mitochondrial orthologs and an unknown
514 mitochondrial role in *P. falciparum*, (ii) proteins without a mitochondrial ortholog but with
515 homology to proteins with a known function and conservation of critical residues that suggest
516 conservation of that function, (iii) proteins with homology to proteins with a known function
517 from which residues known to be essential for catalytic activity or for binding substrates

518 residues have been lost, (iv) new proteins of a known mitochondrial protein family and (v)
519 proteins of a new mitochondrial protein family. Below, we describe several proteins per
520 category, including their domain structure, the conservation of active site residues, their
521 phylogenetic distribution and if available, the experimental confirmation of their
522 mitochondrial localization
523
524 *Proteins with known mitochondrial orthologs and an unknown mitochondrial role in*
525 *P. falciparum*
526 PF3D7_0413500/PfPGM2 (rank 111) contains a PGM domain and appears to be orthologous
527 to the human serine/threonine-protein phosphatase, PGAM5, of which the catalytic residue
528 H105 (61) is conserved (62). PfPGM2 catalyzes the dephosphorylation of phosphorylated
529 sugars and amino acids (62). In human cells, PGAM5 is localized to the mitochondrial outer
530 membrane (63) where it plays a role in apoptosis, mitofission, and mitophagy (reviewed in
531 (64)) among others via interactions with BCL2-like protein 1 (65). A previous study reported
532 an unexpected cytoplasmic localization of this protein in *P. falciparum* and *P. berghei*, though
533 in the absence of brightfield/DIC images or co-localization with marker proteins it is difficult
534 to draw conclusions on the localization (62). Therefore, we tagged the *P. berghei* ortholog
535 PBANKA_0715500 with a 3xHA tag and confirmed co-localization of PbPGM2 with our
536 mitochondrial marker (Fig 3B). We also observed a weak non-mitochondrial background
537 signal, which could indicate a potential dual localization. An important consideration in the
538 interpretation of the previously reported localization of the GFP-tagged *P. berghei* protein is
539 our observation that the use of a big tag such as GFP can interfere with the localization of
540 mitochondrial proteins (Fig S5).
541 Originally, we flagged two additional proteins, PF3D7_0213200 (rank 169) and
542 PF3D7_0611300 (rank 171), as interesting due to potential homology to BCL2, based on
543 sequence similarity and a similar alpha-helical structure predicted with HHpred (36).
544 PF3D7_0213200 was assigned an apicoplast GO-term, but in our analysis is a potential
545 mitochondrial protein. Analysis with TMHMM (66) indicates that it probably is a
546 transmembrane protein with an in-out topology. The protein PF3D7_0611300 already was
547 assigned a mitochondrial GO-term based on a screening in *T. gondii* (67), but was, to our
548 knowledge, not confirmed as mitochondrial localized with experiments characterizing this
549 protein in *P. falciparum.* Although we rejected BCL2 homology for both proteins after a

550 detailed examination of the residues conserved in the BCL2 family, we confirmed the
551 mitochondrial localization of PF3D7_0213200 (PBANKA_0310100 in Fig 3A) and
552 PF3D7_0611300 (PBANKA_0109600 in Fig 3B). Indeed, complexome profiling revealed that
553 the latter is part of the ATP synthase complex (12).
554
555 *Proteins with a non-mitochondrial homolog and conservation of critical residues*
556 PF3D7_0913400 (rank 30) is homologous to the bacterial elongation factor P. Although
557 eukaryotes contain initiation factor 5A as a cytoplasmic elongation factor P homolog, this
558 protein family was until now not observed in mitochondria. PF3D7_0913400 contains both
559 the N-terminal KOW-like domain and the central P/YeiP domain of elongation factor P, but
560 not the C-terminal OB domain. Elongation factor P plays a role in translation elongation and
561 specifically in the translation of stretches of 2 or more prolines (68). A basic residue at position
562 34 of elongation factor P (numbering from the *Escherichia coli* PDB structure 3A5Z_F (69))
563 that after post-translational modification interacts with the CCA-end of the tRNA, is
564 conserved in PF3D7_0913400. Notably, mitochondrial encoded *P. falciparum* cytochrome C
565 oxidase subunit 1 contains a proline pair at position 135-136, making it a potential target for
566 elongation factor P.
567 PF3D7_0812200 (rank 147) is homologous to the bacterial DegP protease family. It
568 contains both a serine endoprotease domain and a PDZ domain. The DegP from *E. coli*
569 functions in acid resistance in the periplasm. It is able to refold after acid stress and
570 subsequently can cleave proteins misfolded due to acid stress (70). The active site residue
571 (S210 in PDB structure 1KY9 (71)) is conserved in PF3D7_0812200, but residues involved in
572 allosteric loop interactions, including R187 are not. Orthologs of PF3D7_0812200 appear,
573 among the eukaryotes, limited to the Apicomplexa and Chromerida.
574 PF3D7_0503900 (rank 246) has a C-terminal dioxygenase domain (E= 1.3E-16). The
575 closest homolog in human is the phytanoyl-CoA dioxygenase that resides in the peroxisome,
576 an organelle absent from *Plasmodium*, where it is involved in α-oxidation of branched-chain
577 fatty acids. Important residues in the active site that are involved in iron binding, H175, H177,
578 H264 (72), are conserved in PF3D7_0503900, suggesting that it is a functional enzyme.
579 Nevertheless, levels of sequence conservation are low (14% identity with the human
580 phytanoyl-CoA dioxygenase) and we did not detect conservation of substrate binding sites

581 with the phytanoyl-CoA dioxygenase. The co-localization analysis suggested a mitochondrial
582 localization for PF3D7_0503900 (PBANKA_1103500 in Fig 3C).
583 PF3D7_1121500 (rank 307) contains a papain-like NLpC/P60 superfamily domain
584 (Peptidase_C92 in PFAM). This protein family contains, among others, cysteine peptidases
585 and amidases. Of the four residues essential for activity in BcPPNE from *Bacillus cereus* (73),
586 the most similar experimentally characterized homolog of PF3D7_1121500, three are
587 conserved (H49, E64, and Y164).
588
589 *Proteins with homology to proteins with a known function from which residues known to be*
590 *essential for catalytic activity or for binding substrates residues have been lost*
591 PF3D7_1004900 (rank 212) is homologous to the protein component of the signal recognition
592 particle (SRP), a ribonucleoprotein that recognizes and targets specific proteins to the ER in
593 eukaryotes. PF3D7_1004900 contains the C-terminal M domain (E= 5.8E-25), but not the
594 other, SRP54 and SRP54_N domains of this protein. In SRP the M domain binds both the SRP
595 RNA and the signal sequence of the target protein. Inspection of individual amino acids did
596 not reveal conservation of RNA binding amino acids, e.g. the arginines in helix M4 of the M
597 domain (74), are not conserved, but did reveal some conservation of hydrophobic amino acids
598 lining the groove in which the signal peptide is located. These hydrophobic amino acids have
599 been implicated in interactions with the signal peptide, specifically V323, I326, L329 in the
600 M1 helix and L418 in the M5 helix (positions relative to *Sulfolobus solfataricus* structure 3KL4
601 (74, 75)), suggesting a possible conservation of interaction of PF3D7_1004900 with a
602 hydrophobic α-helix. We experimentally confirmed the mitochondrial location of
603 PF3D7_1004900 (PBANKA_1203200 in Fig 3A).
604 PF3D7_1246700 (rank 223) is a member of the pyridoxamine 5'-phosphate oxidase
605 (PNPOx) like protein family that, besides pyridoxamine 5'-phosphate oxidase, also contains
606 the general stress protein 26 family. The most similar experimentally characterized protein is
607 general stress protein 26 from *Xanthomonas citri* encoded by gene *Xac2369*. The latter binds
608 FAD and FMN but not pyridoxal 5'-phosphate, indicating that it does not function as a PNPOx
609 (76). Relative to the unpublished structure of a protein from this family from *Nostoc*
610 *punctiforme* (PDB: 2I02), that does contain an FMN, we did not observe conservation of the
611 amino acids like W111, that line the FMN binding pocket, casting doubt of a function of
612 PF3D7_1246700 that includes the binding of FMN.

613

614 *New members of mitochondrial protein families*

615 PF3D7_1417900 (rank 109), PF3D7_1142800 (rank 172) and PF3D7_0927100 (rank 249)
616 potentially contain a CHCH domain. They all have two pairs of cysteines that are very well
617 conserved among their detectable homologs, and the nine amino acids between the cysteines
618 are predicted to form an α-helix (36). Nevertheless, as nothing else but the cysteines appears
619 to be conserved, the E-value with the PFAM CHCH domain entry, or with any known
620 mitochondrial protein, is not significant (E>0.01). Jackhmmer analysis (77) detects homologs
621 among the Apicomplexa, including a homolog in *Cryptosporidium muris*, but not outside this
622 taxon. Known proteins with a CHCH domain are localized to the mitochondrial
623 intermembrane space where they can participate in disulfide bond formation, suggesting that
624 PF3D7_1417900, PF3D7_1142800, and PF3D7_0927100 might also be in the intermembrane
625 space. Complexome profiling revealed that the first two are part of the ATP synthase complex
626 (12). We tested PF3D7_1417900 and PF3D7_0927100 experimentally and confirmed them to
627 be mitochondrial (PBANKA_1024800 and PBANKA_0827900 in Fig 3B respectively). Thus,
628 based on our data and the complexome profiling data all three proteins are confirmed to be
629 mitochondrial, increasing their likelihood that they indeed contain a CHCH domain.

630 *Plasmodium* contains a cytochrome C heme lyase (PF3D7_1224600, rank 129) and a
631 cytochrome C1 heme lyase (PF3D7_1203600, rank 225), which are both essential during
632 blood-stage development and have non-overlapping functions (78). We uncovered a third
633 homolog in this family at rank 190, PF3D7_1121200 (E=0.0096, HHpred (36)). Although the
634 homology covers domains II, III, and IV of heme lyases, of which domain II has been implicated
635 in binding Heme (79), residues that have been implicated in heme binding, like H154
636 (coordinates for human protein (79)) and residues of domain I are not conserved in
637 PF3D7_1121200. The protein also occurs in ciliates and dinoflagellates but has no orthologs
638 outside of those taxa and therewith appears to be an alveolate-specific duplication.

639

640 *A new Apicomplexan mitochondrial protein family*

641 We assessed to what extent *Plasmodium* mitochondrial proteins were part of families with
642 multiple mitochondrial members that resulted from duplications (Supplemental Table 2).
643 Such "intra compartmental protein duplications" have also increased the size of the human
644 mitochondrial proteome (37). Among the mitochondrial homologs, we uncovered a new

645 family in the Apicomplexa that in *Plasmodium* species has four representatives:
646 PF3D7_0821900 (rank 209), PF3D7_1336100 (rank 258), PF3D7_1134400 (rank 280), and
647 PF3D7_1228000 (rank 499). The family is characterized by a well-conserved WPP motif at its
648 N-terminus (Supplemental Fig 7), and we therefore name it the WPP family. Although
649 PF3D7_1228000 does not rank sufficiently high to fall within the 25% cFDR cut-off, all proteins
650 of the family are likely mitochondrial given their overall high ranking and the tendency of
651 members of protein families to be localized in the same cellular location. For the protein
652 family itself, we did not detect homologs outside the Apicomplexa and *Vitrella*. In several
653 species including *Plasmodium coatneyi* (PCOAH_00054040), *Babesia bovis* (BBOV_III010100),
654 and *Theileria orientalis* (MACL_00002472), the ortholog of PF3D7_1228000 is predicted to be
655 fused with an rRNA pseudourydilate synthase hinting at a potential function of this family in
656 rRNA maturation. Nevertheless, there are no experimental data supporting those gene
657 fusions. We unequivocally confirmed the mitochondrial location of PF3D7_0821900
658 (PBANKA_0708900 in Fig 3B) and the co-localization analysis also suggested a mitochondrial
659 localization for PF3D7_1336100 and PF3D7_1134400 (PBANKA_1350000 and
660 PBANKA_0914000 in Fig 3C respectively). PF3D7_1228000 was excluded from the analysis
661 due to low transcript levels.
662
663 **Discussion**
664 In this work, we ranked all nuclear encoded proteins of *P. falciparum* for their likelihood of
665 being mitochondrially targeted and defined the top 445 proteins (cFDR 25%) as likely
666 mitochondrial proteins (Supplemental Table 2). Despite the large variation between the
667 datasets in the coverage of the nuclear encoded proteome and in the ability to identify
668 mitochondrial proteins (Table 1, Fig 1), the probabilistic approach we used was able to
669 integrate the data in an unbiased way by assessing the distribution of gold standard genes
670 within each dataset during the scoring process.
671
672 **Quality of the predictions.** Several observations indicate the high quality of the predictions.
673 Firstly, the 10-fold cross-validation with an area under the curve of 0.959 shows the high
674 predictive value and robustness of the prediction. Secondly, almost half of the predicted
675 proteins have a previous annotation of being mitochondrial, even though most of these were
676 not used in weighing the datasets. Thirdly, of thirteen proteins that we set out to validate

677 experimentally, the seven for which we obtained a high enough expression to convincingly
678 locate them in the cell, were all mitochondrial, while for three others the experimental data
679 was consistent with a mitochondrial localization but equivocal due to low expression levels.
680 Fourth, even though our top 445 contains nine proteins of the negative gold standard that
681 was not manually supervised for all entries, manual examination of those nine showed that
682 the majority are likely mitochondrial after all. Seven of them were detected in the hyperLOPIT
683 study (22) and of those five were annotated as mitochondrial (legend Supplemental Table 2).
684 Finally, several proteins that we predicted to be mitochondrial have, since we made those
685 predictions, been shown to be mitochondrial using other methods. Specifically, complexome
686 profiling of mitochondria-enriched fractions has elucidated the composition of mitochondrial
687 complexes II, III, IV, and V (12). We identify in our top 445 proteins, six out of the seven
688 members of complex II (four of which not in the GS or mitochondrial in Gene Ontology (GO)),
689 ten out of the eleven members of complex III (none of which in the GS, three not in GO),
690 seventeen out of the nineteen members of complex IV (none of which in the GS, thirteen not
691 in GO), and twenty-one of the twenty three members of complex V (fifteen of which not in
692 the GS, nine not in GO, Supplementary Tables 2 and 5). These results reinforce the robustness
693 of the predictions and that the ranked list of the complete nuclear encoded *P. falciparum*
694 proteome, which we name PlasmoMitoCarta, is a useful resource for *Plasmodium*
695 researchers.
696
697 **New mitochondrial biology.** The results of the integration provide ample material for
698 exploring new mitochondrial biology as 185 out of the 445 proteins do not have a detectable
699 mitochondrial ortholog in yeast, human, or plant. Some, however, do have homologs in
700 bacteria that include conservation of critical residues, for example, the elongation factor P
701 whose function, if confirmed, would be unique to mitochondria. They also include paralogs
702 of the known mitochondrial proteins, like the twin Cx9C and heme lyase families, and a new
703 mitochondrial protein family that contains a well-conserved WPP motif. Finally, they include
704 PfPGM2, a protein that does have a human mitochondrial ortholog, but for which the
705 experimental data thus far had however pointed to a cytoplasmic organization (62). The
706 integrated 'omics data and localization with a 3xHA tag do point at a mainly mitochondrial
707 localization for a protein that in human is involved in apoptosis, mitofission, and mitophagy,

708 processes about which little is known in *Plasmodium.* Such proteins provide leads to new
709 mitochondrial biology.
710 As part of this study, we provide two other valuable resources. First, we included the
711 ortholog information gathered for this study that can improve the annotation of *P. falciparum*
712 proteins and provide researchers with additional hints for the function of a protein
713 (Supplemental Table 2). In addition, the WICCA tool that we developed for this study is
714 available online. It allows any researcher to assess co-expression of *P. falciparum* genes with
715 a set of genes of four *Plasmodium* species across data from 83 microarray expression
716 experiments, by weighing the datasets for their relevance to the set of genes. This can aid in
717 the discovery of novel members of known pathways and molecular systems.
718
719 **Size of the mitochondrial proteome.** We predict 445 mitochondrial proteins at the cut off,
720 set at cFDR 25%. When using the posterior probabilities (2^mitochondrial score) of all *P.*
721 *falciparum* genes to calculate the estimated number (E) of mitochondrial proteins by
722 summing them all up (E=sum($P_{posterior}$/($P_{posterior}$+1), see ref (80) for a detailed explanation), we
723 predicted the total size of the mitochondrial proteome to be 454 proteins. It therefore would
724 appear that our prior odds of ~10%, which affects the total size calculation, overestimated
725 the number of mitochondrial proteins. One can in principle lower the prior odds (see ref (80))
726 to 8.3% (412 proteins) such that the number is consistent with the estimated size of the
727 proteome based on all data. However, the presence of five likely mitochondrial proteins in
728 our gold negative sets leads to an underestimate of the size of the mitochondrial proteome.
729 We decided not to manually weed out these inconsistencies, as this cannot be done
730 systematically for all datasets, and applying this only to the apparent inconsistencies would
731 result in circular arguments.
732 The predicted total size of 454 proteins is relatively small compared to the
733 mitochondrial proteomes in human, plant, and yeast that consist of at least 900 proteins (5,
734 39, 40). However, small mitochondrial proteomes have been reported before, for example,
735 573 proteins in the distantly related ciliate *Tetrahymena thermophila* (81). In agreement with
736 our findings, a recent study on the mitochondrial proteome of *T. gondii* that was not included
737 in our integration reports a proteome of 421 proteins (16).
738

739 In conclusion, we combined the information of eight datasets with the curated gold standards
740 to rank all nuclear encoded *P. falciparum* proteins on their likelihood of being mitochondrial
741 in PlasmoMitoCarta. The value of this ranked proteome is shown by *in vivo* validation of top
742 scoring proteins in the closely related species *P. berghei* and will provide an important
743 resource for future investigations.
744

745 **Funding**
746 SLvE and CB were supported by PhD fellowships from the Radboud Institute for Molecular
747 Life Sciences, Radboudumc (Radboudumc JO ronde 2014 and #19-015a). NIP was supported
748 by a Marie-Slodowska Curie grant (790085), and TWAK and LEdV by the Netherlands
749 Organisation for Scientific Research (NWO-VIDI 864.13.009). BED was supported by the
750 Netherlands Organization for Scientific Research (NWO) Vidi grant 864.14.004 and European
751 Research Council (ERC) Consolidator grant 865694: DiversiPHI. KB was supported by a
752 Leverhulme Trust and Isaac Newton Trust Fellowship (ECF-2015-562), and RFW by the
753 Wellcome Trust (214298/Z/18/Z). JvS was funded by the by a grant from the Dutch
754 Organisation for Health Research and Development (ZON-MW TOP grant number 91217009).
755

756 **Competing interests**
757 The authors declare no competing interests.
758
759 **REFERENCES**
760

761 1. Gray MW, Burger G, Lang BF. Mitochondrial evolution. Science (New York, NY).
762 1999;283(5407):1476-81.
763 2. Burger G, Gray MW, Forget L, Lang BF. Strikingly bacteria-like and gene-rich
764 mitochondrial genomes throughout jakobid protists. Genome biology and evolution.
765 2013;5(2):418-38.
766 3. Wilson RJ, Williamson DH. Extrachromosomal DNA in the Apicomplexa. Microbiology
767 and molecular biology reviews : MMBR. 1997;61(1):1-16.
768 4. Makiuchi T, Nozaki T. Highly divergent mitochondrion-related organelles in
769 anaerobic parasitic protozoa. Biochimie. 2014;100:3-17.
770 5. Calvo SE, Clauser KR, Mootha VK. MitoCarta2.0: an updated inventory of mammalian
771 mitochondrial proteins. Nucleic acids research. 2016;44(D1):D1251-7.
772 6. Jedelský PL, Doležal P, Rada P, Pyrih J, Smíd O, Hrdý I, et al. The minimal proteome in
773 the reduced mitochondrion of the parasitic protist Giardia intestinalis. PloS one.
774 2011;6(2):e17285.
775 7. van Dooren GG, Stimmler LM, McFadden GI. Metabolic maps and functions of the
776 Plasmodium mitochondrion. FEMS microbiology reviews. 2006;30(4):596-630.

8. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, et al. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res. 2009;37(Database issue):D539-43.
9. Goodman CD, Siregar JE, Mollard V, Vega-Rodríguez J, Syafruddin D, Matsuoka H, et al. Parasites resistant to the antimalarial atovaquone fail to transmit by mosquitoes. Science (New York, NY). 2016;352(6283):349-53.
10. Sturm A, Mollard V, Cozijnsen A, Goodman CD, McFadden GI. Mitochondrial ATP synthase is dispensable in blood-stage Plasmodium berghei rodent malaria but essential in the mosquito phase. Proceedings of the National Academy of Sciences of the United States 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. Mohammad S, Mohammad A, Ramachandran R, Habib S. [Fe-S] biogenesis and unusual assembly of the ISC scaffold complex in the Plasmodium falciparum mitochondrion. Mol Microbiol. 2021.
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. 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.
16. 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.
17. Matz JM, Goosmann C, Matuschewski K, Kooij TWA. An Unusual Prohibitin Regulates Malaria Parasite Mitochondrial Membrane Potential. Cell reports. 2018;23(3):756-67.
18. 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.
19. 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.
20. 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.
21. 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.
22. 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.
23. 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. Sci Rep. 2018;8(1):410.

822 24. Meerstein-Kessel L, Venhuizen J, Garza D, Proellochs NI, Vos EJ, Obiero JM, et al.
823 Novel insights from the Plasmodium falciparum sporozoite-specific proteome by
824 probabilistic integration of 26 studies. PLoS Comput Biol. 2021;17(4):e1008067.
825 25. Woodcroft BJ, McMillan PJ, Dekiwadia C, Tilley L, Ralph SA. Determination of protein
826 subcellular localization in apicomplexan parasites. Trends in parasitology. 2012;28(12):546-
827 54.
828 26. Günther S, Wallace L, Patzewitz EM, McMillan PJ, Storm J, Wrenger C, et al.
829 Apicoplast lipoic acid protein ligase B is not essential for Plasmodium falciparum. PLoS
830 pathogens. 2007;3(12):e189.
831 27. Read M, Müller IB, Mitchell SL, Sims PF, Hyde JE. Dynamic subcellular localization of
832 isoforms of the folate pathway enzyme serine hydroxymethyltransferase (SHMT) through
833 the erythrocytic cycle of Plasmodium falciparum. Malaria journal. 2010;9:351.
834 28. Boucher MJ, Ghosh S, Zhang L, Lal A, Jang SW, Ju A, et al. Integrative proteomics and
835 bioinformatic prediction enable a high-confidence apicoplast proteome in malaria parasites.
836 PLoS Biol. 2018;16(9):e2005895.
837 29. Li Y, Calvo SE, Gutman R, Liu JS, Mootha VK. Expansion of biological pathways based
838 on evolutionary inference. Cell. 2014;158(1):213-25.
839 30. Verma R, Varshney GC, Raghava GP. Prediction of mitochondrial proteins of malaria
840 parasite using split amino acid composition and PSSM profile. Amino acids. 2010;39(1):101-
841 10.
842 31. Baughman JM, Nilsson R, Gohil VM, Arlow DH, Gauhar Z, Mootha VK. A
843 computational screen for regulators of oxidative phosphorylation implicates SLIRP in
844 mitochondrial RNA homeostasis. PLoS genetics. 2009;5(8):e1000590.
845 32. Szklarczyk R, Megchelenbrink W, Cizek P, Ledent M, Velemans G, Szklarczyk D, et al.
846 WeGET: predicting new genes for molecular systems by weighted co-expression. Nucleic
847 acids research. 2016;44(D1):D567-73.
848 33. Mulvey CM, Breckels LM, Geladaki A, Britovšek NK, Nightingale DJH, Christoforou A,
849 et al. Using hyperLOPIT to perform high-resolution mapping of the spatial proteome. Nature
850 protocols. 2017;12(6):1110-35.
851 34. Christoforou A, Mulvey CM, Breckels LM, Geladaki A, Hurrell T, Hayward PC, et al. A
852 draft map of the mouse pluripotent stem cell spatial proteome. Nature communications.
853 2016;7:8992.
854 35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search
855 tool. Journal of molecular biology. 1990;215(3):403-10.
856 36. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, et al. A Completely
857 Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. Journal of
858 molecular biology. 2018;430(15):2237-43.
859 37. Szklarczyk R, Huynen MA. Expansion of the human mitochondrial proteome by intra-
860 and inter-compartmental protein duplication. Genome biology. 2009;10(11):R135.
861 38. Li L, Stoeckert CJ, Jr., Roos DS. OrthoMCL: identification of ortholog groups for
862 eukaryotic genomes. Genome research. 2003;13(9):2178-89.
863 39. Rao RS, Salvato F, Thal B, Eubel H, Thelen JJ, Møller IM. The proteome of higher plant
864 mitochondria. Mitochondrion. 2017;33:22-37.
865 40. Morgenstern M, Stiller SB, Lübbert P, Peikert CD, Dannenmaier S, Drepper F, et al.
866 Definition of a High-Confidence Mitochondrial Proteome at Quantitative Scale. Cell reports.
867 2017;19(13):2836-52.

868 41. Putignani L, Tait A, Smith HV, Horner D, Tovar J, Tetley L, et al. Characterization of a
869 mitochondrion-like organelle in Cryptosporidium parvum. Parasitology. 2004;129(Pt 1):1-18.
870 42. Pryszcz LP, Huerta-Cepas J, Gabaldón T. MetaPhOrs: orthology and paralogy
871 predictions from multiple phylogenetic evidence using a consistency-based confidence
872 score. Nucleic acids research. 2011;39(5):e32.
873 43. Bender A, van Dooren GG, Ralph SA, McFadden GI, Schneider G. Properties and
874 prediction of mitochondrial transit peptides from Plasmodium falciparum. Molecular and
875 biochemical parasitology. 2003;132(2):59-66.
876 44. Alberts BJ, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. Molecular Biology of the Cell
877 New York: Garland Science; 2002 [4th edition:[The Mitochondrion]. Available from:
878 https://www.ncbi.nlm.nih.gov/books/NBK26894/.
879 45. Kozlowski LP. Proteome-pI: proteome isoelectric point database. Nucleic acids
880 research. 2017;45(D1):D1112-d6.
881 46. Patrickios CS. Polypeptide Amino Acid Composition and Isoelectric Point: 1. A Closed-
882 Form Approximation. Journal of Colloid and Interface Science. 1995;175(1):256-60.
883 47. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al.
884 Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids
885 Res. 2012;40(Database issue):D700-5.
886 48. Barylyuk K, Koreny L, Ke H, Butterworth S, Crook OM, Lassadi I, et al. A
887 Comprehensive Subcellular Atlas of the Toxoplasma Proteome via hyperLOPIT Provides
888 Spatial Context for Protein Functions. Cell Host Microbe. 2020;28(5):752-66 e9.
889 49. Otto TD, Böhme U, Jackson AP, Hunt M, Franke-Fayard B, Hoeijmakers WA, et al. A
890 comprehensive evaluation of rodent malaria parasite genomes and gene expression. BMC
891 biology. 2014;12:86.
892 50. Matz JM, Kooij TW. Towards genome-wide experimental genetics in the in vivo
893 malaria model parasite Plasmodium berghei. Pathogens and global health. 2015;109(2):46-
894 60.
895 51. Deligianni E, Morgan RN, Bertuccini L, Kooij TW, Laforge A, Nahar C, et al. Critical role
896 for a stage-specific actin in male exflagellation of the malaria parasite. Cellular microbiology.
897 2011;13(11):1714-30.
898 52. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an
899 open-source platform for biological-image analysis. Nature methods. 2012;9(7):676-82.
900 53. R Core Team. R: A Language and Environment for Statistical Computing. Vienna,
901 Austria: R Foundation for Statistical Computing; 2015.
902 54. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, et al. gplots:
903 Various R Programming Tools for Plotting Data. 2016.
904 55. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York;
905 2016.
906 56. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier
907 performance in R. Bioinformatics (Oxford, England). 2005;21(20):3940-1.
908 57. Wickham H. scales: Scale Functions for Visualization. 2017.
909 58. Wickham H. Reshaping data with the reshape package. Journal of Statistical
910 Software. 2007;21(12).
911 59. Gupta A, Shah P, Haider A, Gupta K, Siddiqi MI, Ralph SA, et al. Reduced ribosomes of
912 the apicoplast and mitochondrion of Plasmodium spp. and predicted interactions with
913 antibiotics. Open Biol. 2014;4(5):140045.

914 60. Foth BJ, Stimmler LM, Handman E, Crabb BS, Hodder AN, McFadden GI. The malaria
915 parasite Plasmodium falciparum has only one pyruvate dehydrogenase complex, which is
916 located in the apicoplast. Mol Microbiol. 2005;55(1):39-53.
917 61. Takeda K, Komuro Y, Hayakawa T, Oguchi H, Ishida Y, Murakami S, et al.
918 Mitochondrial phosphoglycerate mutase 5 uses alternate catalytic activity as a protein
919 serine/threonine phosphatase to activate ASK1. Proceedings of the National Academy of
920 Sciences of the United States of America. 2009;106(30):12301-5.
921 62. Hills T, Srivastava A, Ayi K, Wernimont AK, Kain K, Waters AP, et al. Characterization
922 of a new phosphatase from Plasmodium. Mol Biochem Parasitol. 2011;179(2):69-79.
923 63. Lo SC, Hannink M. PGAM5 tethers a ternary complex containing Keap1 and Nrf2 to
924 mitochondria. Experimental cell research. 2008;314(8):1789-803.
925 64. Ma K, Chen G, Li W, Kepp O, Zhu Y, Chen Q. Mitophagy, Mitochondrial Homeostasis,
926 and Cell Fate. Frontiers in cell and developmental biology. 2020;8:467.
927 65. Lo SC, Hannink M. PGAM5, a Bcl-XL-interacting protein, is a novel substrate for the
928 redox-regulated Keap1-dependent ubiquitin ligase complex. The Journal of biological
929 chemistry. 2006;281(49):37893-903.
930 66. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane
931 protein topology with a hidden Markov model: application to complete genomes. Journal of
932 molecular biology. 2001;305(3):567-80.
933 67. Sidik SM, Huet D, Ganesan SM, Huynh MH, Wang T, Nasamu AS, et al. A Genome-
934 wide CRISPR Screen in Toxoplasma Identifies Essential Apicomplexan Genes. Cell.
935 2016;166(6):1423-35.e12.
936 68. Ude S, Lassak J, Starosta AL, Kraxenberger T, Wilson DN, Jung K. Translation
937 elongation factor EF-P alleviates ribosome stalling at polyproline stretches. Science (New
938 York, NY). 2013;339(6115):82-5.
939 69. Yanagisawa T, Sumida T, Ishii R, Takemoto C, Yokoyama S. A paralog of lysyl-tRNA
940 synthetase aminoacylates a conserved lysine residue in translation elongation factor P.
941 Nature structural & molecular biology. 2010;17(9):1136-43.
942 70. Fu X, Wang Y, Shao H, Ma J, Song X, Zhang M, et al. DegP functions as a critical
943 protease for bacterial acid resistance. The FEBS journal. 2018;285(18):3525-38.
944 71. Krojer T, Garrido-Franco M, Huber R, Ehrmann M, Clausen T. Crystal structure of
945 DegP (HtrA) reveals a new protease-chaperone machine. Nature. 2002;416(6879):455-9.
946 72. McDonough MA, Kavanagh KL, Butler D, Searls T, Oppermann U, Schofield CJ.
947 Structure of human phytanoyl-CoA 2-hydroxylase identifies molecular mechanisms of
948 Refsum disease. The Journal of biological chemistry. 2005;280(49):41101-10.
949 73. Xu Q, Rawlings ND, Chiu HJ, Jaroszewski L, Klock HE, Knuth MW, et al. Structural
950 analysis of papain-like NlpC/P60 superfamily enzymes with a circularly permuted topology
951 reveals potential lipid binding sites. PloS one. 2011;6(7):e22013.
952 74. Mary C, Scherrer A, Huck L, Lakkaraju AK, Thomas Y, Johnson AE, et al. Residues in
953 SRP9/14 essential for elongation arrest activity of the signal recognition particle define a
954 positively charged functional domain on one side of the protein. RNA (New York, NY).
955 2010;16(5):969-79.
956 75. Janda CY, Li J, Oubridge C, Hernández H, Robinson CV, Nagai K. Recognition of a
957 signal peptide by the signal recognition particle. Nature. 2010;465(7297):507-10.
958 76. Hilario E, Li Y, Niks D, Fan L. The structure of a Xanthomonas general stress protein
959 involved in citrus canker reveals its flavin-binding property. Acta crystallographica Section D,
960 Biological crystallography. 2012;68(Pt 7):846-53.

961 77. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018
962 update. Nucleic acids research. 2018;46(W1):W200-w4.
963 78. Posayapisit N, Songsungthong W, Koonyosying P, Falade MO, Uthaipibull C,
964 Yuthavong Y, et al. Cytochrome c and c1 heme lyases are essential in Plasmodium berghei.
965 Molecular and biochemical parasitology. 2016;210(1-2):32-6.
966 79. Babbitt SE, San Francisco B, Bretsnyder EC, Kranz RG. Conserved residues of the
967 human mitochondrial holocytochrome c synthase mediate interactions with heme.
968 Biochemistry. 2014;53(32):5261-71.
969 80. van Dam TJP, Kennedy J, van der Lee R, de Vrieze E, Wunderlich KA, Rix S, et al.
970 CiliaCarta: An integrated and validated compendium of ciliary genes. PLoS One.
971 2019;14(5):e0216705.
972 81. Smith DG, Gawryluk RM, Spencer DF, Pearlman RE, Siu KW, Gray MW. Exploring the
973 mitochondrial proteome of the ciliate protozoon Tetrahymena thermophila: direct analysis
974 by tandem mass spectrometry. Journal of molecular biology. 2007;374(3):837-63.
975