The Journal of Infectious Diseases

Differential protein expression profiles between Plasmodium falciparum parasites isolated from subjects presenting with pregnancy-associated malaria and uncomplicated malaria in Benin

--Manuscript Draft--

Manuscript Number:	
Full Title:	Differential protein expression profiles between Plasmodium falciparum parasites isolated from subjects presenting with pregnancy-associated malaria and uncomplicated malaria in Benin
Short Title:	P. falciparum protein expression during PAM
Article Type:	Major Article
Section/Category:	Parasites
Keywords:	mass spectrometry, Plasmodium falciparum, pregnancy-associated malaria, field isolate, protein identification, protein abundance.
Corresponding Author:	Philippe Deloron, MD, PhD Research Institute for Developmemt Paris, IdF FRANCE
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Research Institute for Developmemt
Corresponding Author's Secondary Institution:	
First Author:	Gwladys Bertin
First Author Secondary Information:	
Order of Authors:	Gwladys Bertin
	Audrey Sabbagh
	françois Guillonneau
	Sayeh Jafari-Guemouri
	Sem Ezimegnon
	Christian Federici
	Benjamin Houkpatin
	Nadine Fievet
	Philippe Deloron, MD, PhD
Order of Authors Secondary Information:	
Manuscript Region of Origin:	FRANCE
Abstract:	Background: Plasmodium falciparum is responsible of severe malaria, including pregnancy-associated malaria (PAM). During intra-erythrocytic maturation, the infected erythrocyte (iE) membrane is modified by insertion of parasite-derived proteins, mainly consisting of variant surface antigens, such as PfEMP1. Methods: To identify new PAM-specific parasite membrane proteins, we conducted a mass spectrometry-based proteomic study and compared the protein expression profiles between 10 PAM and 10 uncomplicated malaria (UM) samples. Results: We focused on the 454/1139 membrane-associated and hypothetical proteins for comparative analysis. Filter-based feature selection methods combined with supervised data analysis identified a subset of 53 proteins that distinguished PAM and UM samples. Up to 19/20 samples were correctly assigned to their respective clinical group. A hierarchical clustering analysis of these 53 proteins based on the similarity of

Powered by Editorial Manager® and Preprint Manager® from Aries Systems Corporation

their expression profiles revealed two main clusters of 40 and 13 proteins, respectively under- or over-expressed in PAM.
Conclusion: VAR2CSA is identified and associated to PAM, validating our experimental approach. Other PAM-predictive proteins included PFI1785w, PF14_0018, PFB0115w, PFF0325c, and PFA_0410w. These proteomics data demonstrate the involvement of selected proteins in the pathophysiology of PAM, providing new insights for the definition of potential new targets for a vaccine against PAM.

Suggested Reviewers:	Stephen Rogerson sroger@unimelb.edu.au
	Laurence Florens laf@stowers.org
	Christophe Rogier christophe.rogier@gmail.com
	Christopher King cxk21@case.edu
	Peter Kremsner peter.kremsner@uni-tuebingen.de

Powered by Editorial Manager® and Preprint Manager® from Aries Systems Corporation

Martin Hirsch, M.D.
The Journal of Infectious Diseases
225 Friend Street, Suite 401
Boston, MA 02114
USA

Title: Differential protein expression profiles between *Plasmodium falciparum* parasites isolated from subjects presenting with pregnancy-associated malaria and uncomplicated malaria in Benin

Authors: GI. Bertin, A. Sabbagh, F. Guillonneau, S. Jafari-Guemouri, S. Ezimegnon, C. Federici, B. Hounkpatin, N. Fievet, P. Deloron

Dear Editor,

We are pleased to submit a manuscript entitled "Differential protein expression profiles between pregnancy associated malaria and uncomplicated malaria isolates in Benin" for publication in *Journal of Infectious Diseases*.

In this work, we attempted to identify new PAM-specific parasite membrane proteins. To this end, we conducted a mass spectrometry-based proteomic study and compared the composition of the knob structures of *Plasmodium falciparum* isolates associated with pregnancy-associated malaria (PAM) and uncomplicated malaria (UM). Indeed, during the intra-erythrocytic maturation of *P. falciparum*, parasite-derived proteins are successively expressed, exported and presented at the surface of the red blood cell membrane, that allow the export of variant surface antigens (VSAs). In the case of PAM, placental infection is related to the adherence of infected erythrocytes to the placental trophoblast, a phenomenon known to be dependent on the expression of the VAR2CSA variant of VSAs.
We identified by the combination of Mascot and Sequest search algorithms 1139 plasmodial proteins by LC-MS/MS, of which 155 membrane-associated proteins and 299 hypothetical proteins were selected for comparative analysis of clinical groups using a semi-quantitative approach based on emPAI values. Filter-based feature selection methods combined with supervised data analysis identified a subset of 53 proteins that distinguished PAM and UM samples with high classification accuracy. Hierarchical clustering analysis of the expression profiles of these 53 proteins revealed two main clusters of 40 and 13 proteins under-expressed and overexpressed in PAM, respectively.
VAR2CSA has also been identified as the predictive protein of the PAM group, validating our experimental approach. Five other proteins were predictive of PAM parasites: PFI1785w; PF14_0018; PFB0115w; PFA_0410w, and PFF0325c. These novel proteomics data demonstrate the importance of these proteins, and provide new targets for a vaccine against PAM.

Finally, this manuscript has not been submitted elsewhere. All authors have read and approved the manuscript, and there are no conflicts of interest to disclose.

We suggest the following people as potential referees:

- Stephen Rogerson, PhD, Professor, University of Melbourne, Email: sroger@unimelb.edu.au
- Laurence Florens, PhD, the Stowers Institute for Medical Research, Kansas City, Missouri 64110 Email: laf@stowers.org
- Christophe Rogier, Professor, Institut Pasteur, Email: christophe.rogier@gmail.com
- Christopher King, PhD, Welcome Trust Sanger Institute, Email : cxk21@case.edu
- Peter Kremsner, Professor, University of Edinburgh, Email: peter.kremsner@uni-tuebingen.de

For all these reasons we are confident that you will find this paper of interest for publication in *Journal of Infectious Diseases*.

Sincerely yours,

Further correspondence to:
Philippe Deloron

IRD UMR216, Faculté de Pharmacie,
4 avenue de l'Observatoire, 7
5270 Paris Cedex 06, France.

Tel: + 33 1 53 73 96 22; Fax: + 33 1 53 73 96 17

E-mail: philippe.deloron@ird.fr

*Manuscript
Click here to download Manuscript: JID Manuscrit.docx

1 **Differential protein expression profiles between *Plasmodium falciparum* parasites
2 isolated from subjects presenting with pregnancy-associated malaria and uncomplicated
3 malaria in Benin**

4 **Running title: *P. falciparum* protein expression during PAM**

5 Major Article: 3477 words, 47 references, 4 figures, 3 tables

6 Abstract: 194 words

7 Gwladys I. Bertin [1, 2], Audrey Sabbagh [1, 2], François Guillonneau [2, 3], Sayeh Jafari-Guemouri [1,
8 2], Sem Ezimegnon [4], Christian Federici [2, 3, 3bis], Benjamin Hounkpatin [5], Nadine Fievet [1, 2, 4],
9 Philippe Deloron [1, 2 *]

10 [1] Institut de Recherche pour le Développement (IRD), UMR216 - Mère et enfant face aux
11 infections tropicales, Paris, France.

12 [2] PRES Sorbonne Paris Cité, Université Paris Descartes, Paris, France.

13 [3] 3P5 Proteomics facility, Université Paris Descartes, Paris, France.

14 3bis Inserm, U1016, Institut Cochin, Paris, France

15 [4] Centre d'Étude et de Recherche sur le Paludisme Associé à la Grossesse et l'Enfance
16 (CERPAGE), Cotonou, Bénin.

17 [5] Mother and child hospital (HOMEL), Cotonou, Benin.

18 **Footnotes:**

19 - The authors have no conflict of interest.
20 - The works was supported by IRD and Labex GR-Ex.
21 - Reprints and correspondance: Philippe Deloron, UMR216UPD-IRD, Faculté
22 des sciences biologiques et pharmaceutiques; 4, avenue de l'observatoire,
23 75006 Paris France. Phone: + 33 1 53 73 15 27; Fax: + 33 1 53 73 96 17. E-
24 mail: philippe.deloron@ird.fr

25 **Keywords:** mass spectrometry, *Plasmodium falciparum*, pregnancy-associated malaria, field

26 isolate, protein identification, protein abundance.

Summary

Background: *Plasmodium falciparum* is responsible of severe malaria, including pregnancy-associated malaria (PAM). During intra-erythrocytic maturation, the infected erythrocyte (iE) membrane is modified by insertion of parasite-derived proteins, mainly consisting of variant surface antigens, such as PfEMP1.

Methods: To identify new PAM-specific parasite membrane proteins, we conducted a mass spectrometry-based proteomic study and compared the protein expression profiles between 10 PAM and 10 uncomplicated malaria (UM) samples.

Results: We focused on the 454/1139 membrane-associated and hypothetical proteins for comparative analysis. Filter-based feature selection methods combined with supervised data analysis identified a subset of 53 proteins that distinguished PAM and UM samples. Up to 19/20 samples were correctly assigned to their respective clinical group. A hierarchical clustering analysis of these 53 proteins based on the similarity of their expression profiles revealed two main clusters of 40 and 13 proteins, respectively under- or over-expressed in PAM.

Conclusion: VAR2CSA is identified and associated to PAM, validating our experimental approach. Other PAM-predictive proteins included PFI1785w, PF14_0018, PFB0115w, PFF0325c, and PFA_0410w. These proteomics data demonstrate the involvement of selected proteins in the pathophysiology of PAM, providing new insights for the definition of potential new targets for a vaccine against PAM.

48 Introduction

49 Within the erythrocyte, *Plasmodium falciparum* reshapes the host cell, inducing knobs
50 formation. These knobs are responsible for alterations in morphology and deformability of the
51 infected erythrocyte (iE) [1, 2], that increase its membrane rigidity, and prevent the passage of
52 iEs through small capillary blood vessels from the spleen [1]. The major protein present in
53 knobs is KAHRP (knob-associated histidin rich protein) [3], involved in the export of variant
54 surface antigens (VSAs), such as *Plasmodium falciparum* erythrocyte membrane protein-1
55 (PfEMP1). Another major knob protein is MESA (mature parasite-infected erythrocyte
56 surface antigen), involved in destabilizing the erythrocyte membrane skeleton [4, 5].

57 VSAs play an essential role in the host-parasite interface during the intra-erythrocytic cycle
58 [6, 7]. Among VSAs, the two major families of proteins are RIFIN, and PfEMP1. These
59 proteins are coded by multi-gene families, the *repetitive interspersed family* (*rif*) and *var*
60 genes. *Rif* genes are abundant with 200 copies in the genome, coding for the trans-membrane
61 RIFIN proteins on the erythrocyte surface, targets of the immune response, but their function
62 being yet unknown [8]. Unlike *var* genes, several *rif* genes may be concomitantly expressed
63 on the surface of IEs [9].

64 Sixty *var* genes in each genome encode for PfEMP1 proteins that vary in size, but are
65 generally 250-350 kDa [10]. These PfEMP1 are adhesines capable to adhere to endothelial
66 cells or to circulating blood cells. These mechanisms of adhesion are in direct relation with
67 the gravity of the disease [6]. Within the framework of pregnancy-associated malaria (PAM),
68 a particular variant of PfEMP1, the VAR2CSA protein, is involved in the pathogenesis of the
69 disease. Only parasites expressing VAR2CSA, encoded by the *var2csa* gene [11], do sequester
70 in the placenta, the single site of sequestration of *P. falciparum* iE during pregnancy.

71 Several hypothetical proteins have been identified in transcriptomic studies as over-expressed
72 in PAM samples [12, 13], but few have been validated by localization or proteomic studies.

73 The single study linking transcriptomic and proteomic data from malaria isolates allowed to
74 select a subset of genes or proteins differentially expressed between cerebral malaria and
75 PAM [14]. A proteomic study of iE membrane proteins from placental *P. falciparum* isolates
76 [15] identified proteins with a low score of identification and a reduced number of peptides,
77 probably a consequence of the relative poverty in parasitic proteins. Indeed, parasite proteins
78 are embedded in a mass of proteins from the erythrocyte membrane skeleton that are much
79 more abundant, impairing the identification of parasite proteins by MS.

80 The aim of this work was to compare the expression profiles of membrane proteins involved
81 in knobs structure from isolates associated with two distinct clinical presentations of malaria,
82 PAM and UM, using a proteomic approach.

83

84 Experimental Procedures

85 Ethic statement

86 Ethical clearance was obtained from the Institutional Ethics Committee of the Faculty of
87 Health Science, Abomey-Calavi University. All patients were included after written informed
88 consent from themselves or their guardian. Patients received care and adequate treatment,
89 according to the national malaria program policy.

90 Subjects enrolment and sample collection

91 Patients were enrolled in southern Benin from May to August 2011. Pregnant women
92 presenting with PAM were included in the Mother and Child hospital (Hôpital de la Mère et
93 de l'Enfant Lagune), Cotonou. Uncomplicated malaria (UM) patients were enrolled in the
94 health centre of Come, 70 km from Cotonou.

95 Uncomplicated malaria (UM) was defined as the combination of fever (tympanic temperature
96 ≥37.8°C), confirmed presence of *P. falciparum* in the peripheral blood, and absence of any
97 severity sign, as defined by the World Health Organization [16]. Patients and pregnant women
98 at delivery were screened by rapid diagnostic test for *P. falciparum* (Malaria Quick test,
99 Cypress Diagnostics, Langdorp, Belgium). Blood sample (5 ml) was collected in a vacutainer
100 tube containing EDTA, and a placental blood sample was obtained after delivery. Giemsa-
101 stained thick blood film confirmed *P. falciparum* infection. Thirty-two isolates from UM gave
102 mature parasites, after less than 18h of *in vitro* culture. Fifteen infected placentas were
103 collected and flushed from fresh. The 47 samples (UM and PAM) containing mature parasite
104 forms were depleted from uninfected erythrocytes over a Macs column [17]. Then, samples
105 were lysed according to [15] and stored at -80°C.

106 To prepare trypsin digested peptides, 100 µg of proteins were reduced with 20 mM DTT
107 during 30 min at 56°C, then alkylated with 55 mM of chloroacetamide for 30 min at room
108 temperature. The precipitate was resuspended in digestion buffer (50 mM ammonium

109 bicarbonate, 2 % rapigest, 20 mM DTT and 1 µg/µl of trypsin) and digested overnight at
110 37°C.
111 LC –MS/MS analysis
112 Analyses were performed using an Ultimate 3000 Rapid Separation liquid chromatographic
113 system (Dionex, The Netherlands) coupled to a hybrid Linear Trap Quadrupole-ORBITRAP
114 Velos mass spectrometer (Thermo Fisher Scientific, San José CA). Peptides were separated
115 on a C18 RP analytical column (3 µm particle size, 100 Å pore size, 75 µm i.d., 50 cm length)
116 with a 240-minute gradient from 99 % A (ACN 5 %, formic acid 0.1 % and H_2O 95 %) to 40
117 % B (ACN 80 %, formic acid 0.085 % and H_2O 20 %).
118 Protein identification and compilation of search results
119 All LC-MS/MS results were analyzed using Proteome Discoverer 1.3.0 software
120 (ThermoFisher Scientific, CA, USA) in combination with Mascot [18] and Sequest [19]
121 search algorithms. MS/MS data were searched against a database representing the
122 concatenation of – NCBI human - NCBI *Plasmodium falciparum* - Vardom *var* genes
123 sequences and their respective reverse sequences. Precursor mass tolerance was set to 2 ppm
124 and fragment mass tolerance to 0.45 Da. False positive probability was < 5%. All proteins
125 identified with both database search algorithms (Mascot and Sequest), presenting at least two
126 peptides with a score of identification of proteins ≥ 20 for Mascot, and an Xcorr ≥ 1.5 for
127 Sequest were considered as positive hits.
128 Estimating absolute protein abundance values
129 Protein abundance values were determined using the exponentially modified Protein
130 Abundance Index (emPAI) [20, 21], calculated according to the following formula.
131 $$emPAI = 10^{PAI} - 1 \text{ with } PAI = N_{\text{observed peptides}}/N_{\text{observable peptides}}$$
132 The number of "observable" peptides was calculated from the output of the Protein Digestion
133 Stimulator program (http://omics.pnl.gov/software/ProteinDigestionSimulator.php). Samples

134 with missing emPAI value for a particular protein were assigned half the minimum emPAI
135 value for that protein [22, 23]. Normalization between samples was performed according to
136 the median of each sample. EmPAI values were then log2 transformed.

137 Unsupervised clustering analysis

138 Clustering, or unsupervised learning, aims to group samples on the basis of similarity between
139 the protein expression profiles. Agglomerative hierarchical clustering [24] organizes all data
140 elements into a single tree (or dendrogram) with the highest levels of the tree representing the
141 discovered classes. Samples were clustered using average linkage clustering and with
142 Spearman correlation as similarity metric. We used *k*-means [25, 26] and self-organizing map
143 [27], two non-hierarchical clustering techniques, which partition samples into different
144 clusters without trying to specify the relationship between individual elements. We ran
145 MacQueen algorithm with 50 trials and used Euclidean distance as the similarity metric.
146 Finally, unsupervised principal component analysis (PCA) [28] was applied to the protein
147 data and the top components were used to illustrate the similarity in protein expression
148 profiles among PAM and UM samples.

149 Supervised classification analysis

150 Supervised class prediction analysis was performed using five machine-learning algorithms to
151 identify the best set of predictors (proteins) yielding the most accurate class assignments
152 (PAM *vs.* UM): the classification tree (CART) [29], *k*-nearest neighbors [30], multilayer
153 perceptron neural network [31], radial basis function neural network [32], and support vector
154 machine [33] methods. The simplest tree was selected on the basis of the 1-SE rule i.e. the
155 simplest tree for which the pruning error rate was not significantly higher than the one of
156 optimal tree. We used the GINI index as an indicator of goodness of split in the growing
157 phase, and 35 % of the samples were used as a separate pruning set in the post-pruning
158 process. We chose the heterogeneous Euclidean overlap metric distance in our *k*-nearest

159 neighbors classifier with $k = 3$, and predicted the class by majority vote. The neural
160 architecture parameters of the multilayer perceptron were one hidden layer with 20 neurons,
161 and the learning rate was fixed to 0.15. For the support vector regression model with a radial
162 basis function, we used the off-line learning of kernels implemented in Tanagra [34] with
163 default parameters. Finally, we used a linear support vector machine which implements Platt's
164 sequential minimal optimization algorithm [35], and trained the vector support classifier using
165 a polynomial kernel of degree 2.

166 Classification accuracy was evaluated by leave-one-out cross-validation (LOOCV). In
167 LOOCV, one sample is excluded from the data set, and the remaining samples are used to
168 build the classifier. Then the classifier is used to predict the class of the left out one, and this
169 is repeated for each sample in the data set.

170 Filter-based methods for feature selection

171 Three filter techniques were applied to select the most relevant features (proteins) that
172 discriminate between PAM and UM samples: Fisher discriminant criterion, Runs test and
173 ReliefF. Fisher filtering is one of the most widely used criteria for supervised feature selection
174 due to its general good performance. It follows univariate Fisher's ANOVA ranking which
175 ranks the input features according to their relevance [36]. Fisher's criterion takes the mean
176 and the within-class scatter of the groups into account to compare the correlation between
177 features and the class label. Runs filtering is a non-parametric test for predictive feature
178 evaluation [37], performing an univariate attribute ranking from the runs test, also known as
179 Wald-Wolfowitz test, that checks a randomness hypothesis for a two-valued data sequence
180 [38]. In both Fisher and Runs filtering, the significance cut-off P-value was 0.05. ReliefF is an
181 extension of the Relief algorithm [39]. We selected as relevant features proteins whose
182 weights were higher than the mean of positive weights. The subsets of features selected by
183 each of the filter-based methods, or by a combination of these, were presented as input to the

184 supervised classification algorithms to evaluate their performance, and the feature subset with
185 the highest classification accuracy was kept as final set.
186 All statistical analyses were performed using GenePattern v3.5
187 (http://www.broadinstitute.org/cancer/software/genepattern/) [40] and Tanagra v1.4
188 (http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html) [34] softwares.
189 Transmembrane domain prediction
190 All sequences of proteins identified by LC-MS/MS were submitted to transmembrane
191 domains prediction algorithms through
192 http://www.bioinformatics.utep.edu/BIMER/tools/transmembrane.html using five algorithms:
193 DAS, HMMTOP, TMpred, TMHMM server2.0, and TopPred 0.01. We considered as
194 "correct" the prediction of a transmembrane domain by at least two algorithms.
195 Protein-protein interaction network.
196 The protein-protein interaction network was built using the Search Tool for the Retrieval of
197 INteracting Genes/proteins (STRING 9.0) database [41] at http://string.embl.de/. Proteins
198 overexpressed in PAM were given as input to the STRING database, as well as four proteins,
199 which the encoding genes were identified as associated to PAM [14]. A search for known and
200 predicted protein-protein interactions was performed using a confidence score of 0.15 and
201 0.40 (low and medium confidence, respectively).
202

203 Results

204 Clinical groups

205 Ten patients with uncomplicated malaria (UM group) and ten delivering women (PAM group)
206 were included. Among UM, 2/10 were female, mean age was 12.5 years, and mean
207 parasitemia was 125,068 parasites/µl. Among pregnant women, 2 were primigravidae, mean
208 age was 25.1 years, and mean parasitemia was 69,750 parasites/µl.

209 Descriptive data

210 Ten parasite samples from each clinical group were analyzed by LC-MS/MS, and proteins
211 were identified using two database search algorithms (Mascot and Sequest). Ninety-four
212 percent (2768/2948) of the proteins were identified by both algorithms (all proteins identified
213 by Mascot were also identified by Sequest, and 180 proteins were unique to Sequest),
214 ensuring the reliability of protein identification. Proteins identified in PAM samples were
215 twice as numerous as in UM (1876 *vs.* 965). This difference is due to the higher number of
216 human proteins in PAM samples, and is likely to arise from differences in sample origin
217 (peripheral blood for UM samples vs. placental blood for PAM samples). The number of
218 *Plasmodium* proteins was similar in PAM and UM samples (762 *vs.* 672). The cellular
219 localization of proteins identified in UM and PAM samples was similar (Figure 1). Membrane
220 proteins represented ~15 % of all identified proteins in both groups; 18 % to 28 % of proteins
221 were not annotated.

222 Human membrane proteins were identified in all samples from both clinical groups,
223 demonstrating the ability to identify membrane proteins (Table 1). Table 1 shows parasite
224 proteins expressed at the membrane of erythrocytes infected by mature parasites (trophozoite
225 and schizont) that were identified in all samples, demonstrating a similar maturation stage.

226 Identification of proteins differentially expressed in the two parasite populations

227 Differences in protein expression profiles between PAM and UM samples were evaluated
228 using a semi-quantitative approach by computing emPAI values for all proteins [20].
229 The twenty samples yielded 1139 *Plasmodium* proteins identified in at least one sample. We
230 focused on 454 of these (40%), identified as either membrane-associated (N = 154) or
231 hypothetical (N = 300) proteins.
232 We first conducted an exploratory analysis by unsupervised clustering methods. None of the
233 clustering methods (hierarchical clustering, k-means and SOM) allowed to partition data into
234 clusters corresponding to the PAM and UM groups (Supplemental Figure 1). This result was
235 expected given the overwhelming number of features (proteins) relative to the number of
236 available samples, leading to the so-called "curse of dimensionality". We combined three
237 filter-based methods to select the subset of proteins that best discriminate PAM from UM
238 samples. From the 454-protein set, Fisher filtering, Runs filtering and ReliefF selected subsets
239 of 61, 24, and 72 proteins, respectively (Table 2a). Protein lists produced by the different
240 methods largely agreed: 72-85 % of proteins were common to two or three lists. We compiled
241 these results to generate a new list composed of proteins selected by at least two filter
242 techniques. This led to a fourth list of 53 proteins.
243 To select the subset of proteins more likely of biological significance, we evaluated the
244 performance of the different feature subsets with supervised classification methods. Since
245 different types of classifiers can respond differently to the same input data, five different
246 classification tools were used, including the classification tree (CART), *k*-nearest neighbors
247 (*k*NN), multilayer perceptron neural network (MPNN), radial basis function neural network
248 (RBFNN), and support vector machine (SVM) methods. We estimated the prediction
249 accuracy of the new list of 53 proteins together with the three protein lists produced by the
250 individual filters; and compared the results with those obtained with the initial set of 454
251 proteins (i.e. before filtering) (Table 2b). Among the different feature subsets, the best

252 classification results were always obtained with the 53-protein consensus list, with a correct
253 classification rate ranging from 85% (3/20 samples misclassified) to 95% (a single sample
254 misclassified). The best classification accuracy was achieved by *k*NN, MPNN and SVM
255 methods, which only misclassified the PAM4 sample as UM. In addition to this error, the
256 CART method misclassified the UM AS3 sample as PAM, and the RBFNN method made a
257 third error by incorrectly assigning UM status to the PAM HM037 sample.
258 Unsupervised hierarchical clustering of the 20 samples based on their expression profiles for
259 the 53 discriminatory proteins confirmed these results by partitioning the samples into two
260 main clusters corresponding to the PAM and UM groups, except for the PAM4 sample which
261 clustered with the UM samples, and the HM037 sample which branched outside the two main
262 clusters. Two main clusters of proteins emerged: the first one includes 40 proteins under-
263 expressed in PAM samples as compared to UM, and the second one includes 13 proteins
264 over-expressed in PAM (Figure 2, Table 3). Similar results were obtained with the two other
265 clustering methods (Figure S2). Similarity in protein expression profiles among the 20
266 samples was graphically summarized in a 2-dimensional scatterplot of the two first principal
267 components of the PCA (Figure 3). The first principal component (PC1), which accounts for
268 41 % of total variation, showed a clear-cut separation between PAM and UM samples. Again,
269 PAM4 was closer to the UM samples than to members of its own clinical group.
270

271 ## Discussion

272 We used a mass-spectrometry-based proteomic approach to compare the membrane and
273 hypothetical proteins of *P. falciparum* parasites from two clinical forms of malaria (PAM *vs.*
274 UM) in Benin. We aimed to identify membrane-associated proteins specifically associated or
275 over-expressed in PAM. Analysis by LC-MS/MS of ten blood samples from each population,
276 and their comparison allowed to identify a total of 1139 *Plasmodium* proteins, of which 454
277 were membrane or hypothetical proteins. The high dimensionality of data, arising from the
278 high number of features (proteins) compared to the relatively low number of samples
279 available, poses a great challenge to identify the proteins that best discriminate between the
280 two classes of parasites. Hence, the use of an efficient feature selection technique seems to be
281 an appropriate approach to reduce dimensionality, and to improve results interpretability.
282 Such methods are increasingly used in microarray and mass spectrometry domains [42, 43].
283 Instead of choosing one feature selection method, we used three different methods based on
284 univariate (Fisher's ANOVA and runs test) or multivariate (ReliefF) filtering. The success of
285 each subset of features as a classifier was evaluated through supervised classification analysis
286 using five different tools, after which the best set was kept. This set included 53 differentially
287 expressed proteins, which discriminated PAM and UM samples with high accuracy.
288 Unsupervised analysis, including clustering and PCA, based on the expression profiles of
289 these 53 proteins, clearly distinguished samples according to their clinical status, except for
290 PAM4 and HM037. Interestingly, the donors of both samples presented a low parasite density
291 (8000 parasites/μl) compared to the others (69,750 parasites/μl). HM037 comes from a
292 multiparous woman with a low parasitemia who delivered a normal weight baby, and an
293 overall low number of membrane proteins were identified (4.4% of membrane proteins
294 against 15% in the other samples). This may explain why PAM-specific proteins, such as

295 VAR2CSA and PFI1785w (see below), were not identified in PAM4, and why HM037 did
296 not cluster with any of the other samples in the hierarchical clustering analysis (Figure 2).
297 Interestingly, among the proteins showing the most contrasting expression pattern between
298 PAM and UM samples, five proteins overexpressed in UM belong to the RIFIN protein
299 family (Table 3). Their alignment revealed a conserved motif typical of the RIFIN A group
300 [44], and the presence of three transmembrane domains. Among the proteins identified as
301 predictive of PAM, the VAR2CSA protein appears through four independent identifications
302 (Figure 2, Table 3). VAR2CSA is specifically expressed by PAM parasites [11] [45], and is
303 currently the leading candidate for an anti-PAM vaccine. It can therefore be considered as a
304 positive control in our analyses. Besides PfEMP1, six other proteins were also identified as
305 predictive of PAM (Table 3). Merozoite Surface Protein-1 (MSP1) is also a target for malaria
306 vaccine because of its involvment in invasion [46]. PFI1785w and PF14_0018 belong to the
307 *Plasmodium* helical interspersed sub-telomeric protein family (PHIST), a family of proteins
308 proposed to serve generally as interaction modules linking parts of the parasite intra-
309 erythrocyte protein network and trafficking to the erythrocyte membrane [15, 47]. The
310 PFI1785w protein is encoded by a gene overexpressed in PAM isolates, as shown in
311 transcriptome studies [12, 13, 15]. A proteomic study also pointed out the higher abundance
312 of PFI1785w protein in PAM isolates from Tanzania, and confirmed its presence by Western-
313 blot [15]. It is noteworthy that the PFI1785w protein is the most correlated with the first axis
314 of the PCA (correlation coefficient $r^2 = 0.73$) that clearly distinguished PAM from UM
315 samples (Supplemental Figure S3). The VAR2CSA protein follows right behind. Although
316 localization and function of PFB0115w are unknown, this protein was suggested to be
317 involved in proteins trafficking within the erythrocyte by alternate export pathways [12].
318 PFB0115w has also been described as overexpressed in PAM parasites in Tanzania, at both
319 the gene and protein levels [14, 15]. Our findings on parasites isolated from pregnant women

320 from West Africa thus corroborate previous observations made with a similar technology on
321 parasites from East Africa. PFA_0410w and PFF0325c are conserved *Plasmodium* proteins
322 with unknown function.
323 Figure 4 shows the network of protein-protein interactions built using the STRING database
324 [41] with the seven proteins identified as predictive of PAM in this study, as well as
325 PFD1140w, PFL0050c, PFE1640w (var1csa) and MAL13P1.470, which the corresponding
326 genes where found overexpressed in PAM in [14]. Interestingly, all proteins were connected
327 in the same network with a score of 0.40. The PF14_0018 appears to be closely associated
328 with AMA1 and PFA_0440w, and PFA_0440w is directly linked to PFF0325c protein.
329 ETRAMP5 protein seems to be essential to link PFA_0410w with the remaining proteins,
330 suggesting its sub-membrane localization. In this network, PFI1785w is a direct partner of
331 VAR2CSA, and is probably extracellular with one transmembrane domain and one PEXEL
332 motif. Conversely, PFB0115w which is associated with PFI1785w is probably a
333 submembrane protein and indirectly linked to VAR2CSA.
334 These proteomics data demonstrate the involvement of selected proteins in the
335 pathophysiology of PAM, providing new insights for the definition of potential new targets
336 for a vaccine against PAM.
337

338 Acknowledgements

339 We are grateful to all pregnant women and patients and their parents for participating to this
340 study. We are thankful to the nurse in HOMEL, for her help in the collection of samples. We
341 thank Virginie Salnot, Marjorie Leduc and Laila Sago for their help in performing and
342 analyzing proteomics data. We thank Célia Dechavanne for her competence in STATA 12
343 software. We thank also Benoît Hareng and Fatma Mounsi for the preparation and extraction
344 of protein samples.

345

346 References

347 1. Deitsch KW, Wellems TE. Membrane modifications in erythrocytes parasitized by
348 Plasmodium falciparum. Mol Biochem Parasitol **1996**; 76:1-10.
349 2. Cooke BM, Mohandas N, Coppel RL. The malaria-infected red blood cell: structural and
350 functional changes. Adv Parasitol **2001**; 50:1-86.
351 3. Kilejian A. Characterization of a protein correlated with the production of knob-like
352 protrusions on membranes of erythrocytes infected with Plasmodium falciparum. Proceedings
353 of the National Academy of Sciences of the United States of America **1979**; 76:4650-3.
354 4. Waller KL, Stubberfield LM, Dubljevic V, et al. Interactions of Plasmodium falciparum
355 erythrocyte membrane protein 3 with the red blood cell membrane skeleton. Biochim Biophys
356 Acta **2007**; 1768:2145-56.
357 5. Pei X, An X, Guo X, Tarnawski M, Coppel R, Mohandas N. Structural and functional
358 studies of interaction between Plasmodium falciparum knob-associated histidine-rich protein
359 (KAHRP) and erythrocyte spectrin. J Biol Chem **2005**; 280:31166-71.
360 6. Coppel RL, Brown GV, Nussenzweig V. Adhesive proteins of the malaria parasite. Curr
361 Opin Microbiol **1998**; 1:472-81.
362 7. Abdel-Latif MS, Khattab A, Lindenthal C, Kremsner PG, Klinkert MQ. Recognition of
363 variant Rifin antigens by human antibodies induced during natural Plasmodium falciparum
364 infections. Infect Immun **2002**; 70:7013-21.
365 8. Kyes S, Horrocks P, Newbold C. Antigenic variation at the infected red cell surface in
366 malaria. Annu Rev Microbiol **2001**; 55:673-707.
367 9. Kyes S, Pinches R, Newbold C. A simple RNA analysis method shows var and rif
368 multigene family expression patterns in Plasmodium falciparum. Mol Biochem Parasitol
369 **2000**; 105:311-5.
370 10. Gardner MJ, Shallom SJ, Carlton JM, et al. Sequence of Plasmodium falciparum
371 chromosomes 2, 10, 11 and 14. Nature **2002**; 419:531-4.
372 11. Salanti A, Staalsoe T, Lavstsen T, et al. Selective upregulation of a single distinctly
373 structured var gene in chondroitin sulphate A-adhering Plasmodium falciparum involved in
374 pregnancy-associated malaria. Mol Microbiol **2003**; 49:179-91.
375 12. Francis SE, Malkov VA, Oleinikov AV, et al. Six genes are preferentially transcribed by
376 the circulating and sequestered forms of Plasmodium falciparum parasites that infect pregnant
377 women. Infect Immun **2007**; 75:4838-50.
378 13. Tuikue Ndam N, Bischoff E, Proux C, et al. Plasmodium falciparum transcriptome
379 analysis reveals pregnancy malaria associated gene expression. PloS one **2008**; 3:e1855.
380 14. Vignali M, Armour CD, Chen J, et al. NSR-seq transcriptional profiling enables
381 identification of a gene signature of Plasmodium falciparum parasites infecting children. J
382 Clin Invest **2011**; 121:1119-29.
383 15. Fried M, Hixson KK, Anderson L, Ogata Y, Mutabingwa TK, Duffy PE. The distinct
384 proteome of placental malaria parasites. Mol Biochem Parasitol **2007**; 155:57-65.
385 16. Organization WH. Severe falciparum malaria. Communicable Diseases Cluster. Trans R
386 Soc Trop Med Hyg 94 Suppl 1:S1-90. **2000**.
387 17. Ribaut C, Berry A, Chevalley S, et al. Concentration and purification by magnetic
388 separation of the erythrocytic stages of all human Plasmodium species. Malaria journal **2008**;
389 7:45.
390 18. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification
391 by searching sequence databases using mass spectrometry data. Electrophoresis **1999**;
392 20:3551-67.

393 19. Eng JK MA, and Yates JR III. An Approach to Correlate Tandem Mass Spectral Data of
394 Peptides with Amino Acid Sequences in a Protein Database. J Am Soc Mass Spectrom **1994**;
395 5:976 - 89.
396 20. Ishihama Y, Oda Y, Tabata T, et al. Exponentially modified protein abundance index
397 (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced
398 peptides per protein. Mol Cell Proteomics **2005**; 4:1265-72.
399 21. Silvestrini F, Lasonder E, Olivieri A, et al. Protein export marks the early phase of
400 gametocytogenesis of the human malaria parasite Plasmodium falciparum. Mol Cell
401 Proteomics **2010**; 9:1437-48.
402 22. Makawita S, Smith C, Batruch I, et al. Integrated proteomic profiling of cell line
403 conditioned media and pancreatic juice for the identification of pancreatic cancer biomarkers.
404 Mol Cell Proteomics **2011**; 10:M111 008599.
405 23. Wu CC, Hsu CW, Chen CD, et al. Candidate serological biomarkers for cancer identified
406 from the secretomes of 23 cancer cell lines and the human protein atlas. Mol Cell Proteomics
407 **2010**; 9:1100-17.
408 24. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-
409 wide expression patterns. Proceedings of the National Academy of Sciences of the United
410 States of America **1998**; 95:14863-8.
411 25. Forgy EW. Cluster analysis of multivariate data: efficiency versus interpretability of
412 classifications. Biometrics **1965**; 21:768-9.
413 26. MacQueen JB. Some Methods for classification and Analysis of Multivariate
414 Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and
415 Probability. University of California Press pp **1967**:281-97.
416 27. Kohonen T. Self-organization and associative memory. Springer-Verlag **1988**.
417 28. Abdi HaW, L. J. . Principal component analysis. Wiley Interdisciplinary Reviews:
418 Computational Statistics **2010**; 2:433-59.
419 29. Breiman L, Friedman, J., Olsen, R., Stone, C. Classification and Regression Trees. .
420 Chapman & Hall **1984**.
421 30. Aha D, Kibler, D., Albert, M. Instance-based learning algorithms. Machine Learning
422 **1991**; 6:37-66.
423 31. Mehrotra K, Mohan, C., Ranka, S. Elements of artificial neural network. MIT Press
424 **1997**:66-87.
425 32. Buhmann MD. Radial Basis Functions: Theory and Implementations. Cambridge
426 University **2003**.
427 33. Cristianini NaS-T, J. An Introduction to Support Vector Machines and other kernel-based
428 learning methods. Cambridge University Press **2000**.
429 34. Rakotomalala R. TANAGRA : un logiciel gratuit pour l'enseignement et la recherche. In
430 Actes de EGC'2005, RNTI-E-3 **2005**; 2:697-702.
431 35. Platt J. Fast Training of Support Vector Machines using Sequential Minimal
432 Optimization. Advances in Kernel Methods - Support Vector Learning, B Schölkopf, C
433 Burges, and A Smola, eds, MIT Press **1998**.
434 36. Fisher RA. The use of multiple measurements in taxonomic problems. . Ann Eugenics
435 **1936**; 7:179-88.
436 37. Rakotomalala R. Graphes d'induction. PhD thesis (http://eric.univ-
437 lyon2.fr/~ricco/doc/Graphes_Induction_These_Rakotomalala_1997.pdf) **1997**.
438 38. Wald AaW, J. . On a test whether two samples are from the same population. The Annals
439 of Mathematical Statistics **1940**; 11:147-62.
440 39. Robnik-Sikonja M, Kononenko, I. Theoretical and empirical analysis of ReliefF and
441 RReliefF. . Mach Learn **2003**; 53:23-69.

442 40. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet
443 **2006**; 38:500-1.
444 41. Franceschini A, Szklarczyk D, Frankild S, et al. STRING v9.1: protein-protein interaction
445 networks, with increased coverage and integration. Nucleic Acids Res **2013**; 41:D808-15.
446 42. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics.
447 Bioinformatics **2007**; 23:2507-17.
448 43. Datta S, Pihur V. Feature selection and machine learning with mass spectrometry data.
449 Methods Mol Biol **2010**; 593:205-29.
450 44. Joannin N, Abhiman S, Sonnhammer EL, Wahlgren M. Sub-grouping and sub-
451 functionalization of the RIFIN multi-copy protein family. BMC Genomics **2008**; 9:19.
452 45. Tuikue Ndam NG, Salanti A, Bertin G, et al. High level of var2csa transcription by
453 Plasmodium falciparum isolated from the placenta. The Journal of infectious diseases **2005**;
454 192:331-5.
455 46. Moss DK, Remarque EJ, Faber BW, et al. Plasmodium falciparum 19-kilodalton
456 merozoite surface protein 1 (MSP1)-specific antibodies that interfere with parasite growth in
457 vitro can inhibit MSP1 processing, merozoite invasion, and intracellular parasite
458 development. Infect Immun **2012**; 80:1280-7.
459 47. Kilili GK, LaCount DJ. An erythrocyte cytoskeleton-binding motif in exported
460 Plasmodium falciparum proteins. Eukaryot Cell **2011**; 10:1439-47.

461

462 Figure legends

463 **Figure 1: Proportion of proteins identified according to their cellular localization, by**
464 **clinical group**

465 Each color corresponds to one category of proteins: cytoplasmic (red), membrane (green),
466 nuclear (violet), cytosolic (blue), others (orange), and not annotated (dark blue). A/ proteins
467 identified in uncomplicated malaria (UM) samples and B/ proteins identified from placental
468 blood in pregnancy-associated malaria (PAM) samples.

469 **Figure 2: Hierarchical clustering analysis of the 20 samples based on the expression**
470 **profile of the 53 discriminatory proteins.** Both samples and proteins were clustered using
471 average linkage clustering and with Spearman correlation as similarity metric. The samples
472 are shown horizontally (columns), the proteins vertically (rows). The dendrograms represent
473 the distances between the clusters. In the heat map of protein expression patterns, expression
474 levels are represented in the color scale from blue (low expression) to red (high expression).

475 **Figure 3: Principal component analysis based on the expression profiles of the 53**
476 **discriminatory proteins.** Blue and red dots represent uncomplicated malaria (UM) and
477 pregnancy-associated malaria (PAM) samples, respectively. Each axis represents a principal
478 component (PC1 and PC2) with the percentage of the total variance. The next two
479 components (PC3 and PC4) explain only 9.6 % and 8.4 % of total variance, respectively.

480 **Figure 4**: **STRING 9.0 generated network of PAM-associated proteins.** The map shows
481 the confidence view of the network with stronger associations being represented by thicker
482 (blue) lines. The proteins are identified by their gene names located near each sphere. (1)
483 shows the proteins identified by LC-MS/MS analysis. (2) indicates the proteins, which the
484 encoding genes were identified as associated to PAM by Vignali [14]. (3) indicates the PAM-
485 associated proteins identified by Fried [15].

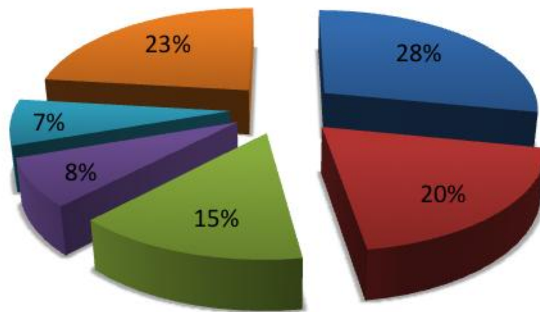
486

487

Figure
Click here to download Figure: Figure.doc

Figure 1:

A/ UM



B/ PAM

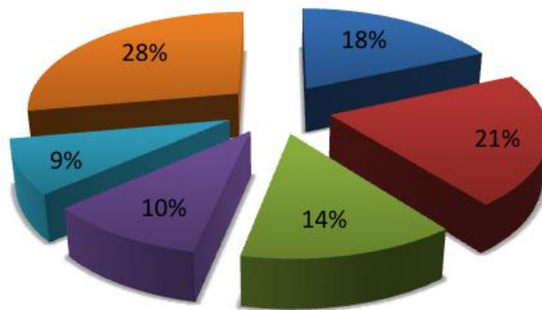


Figure 2:

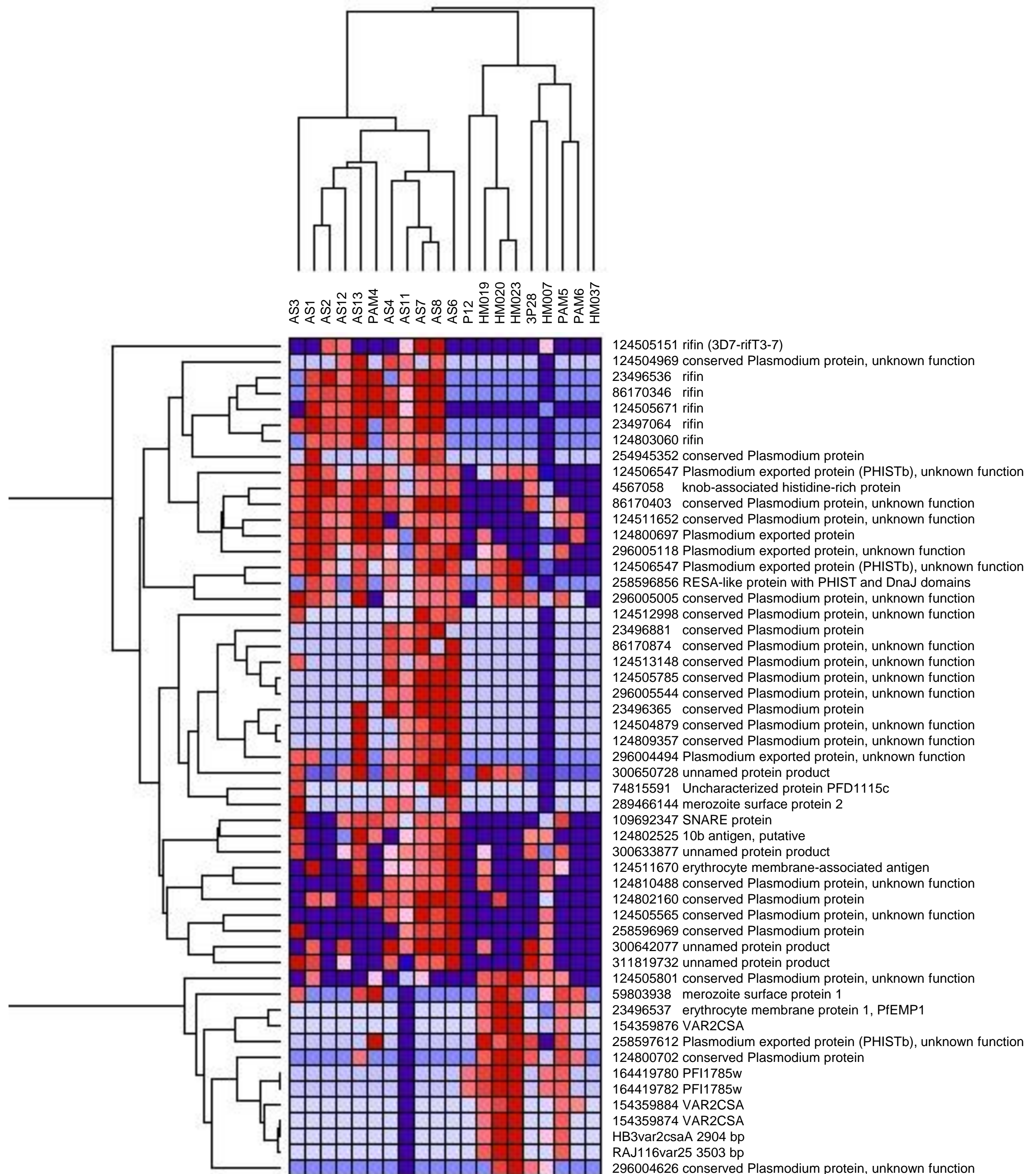


Figure 3:

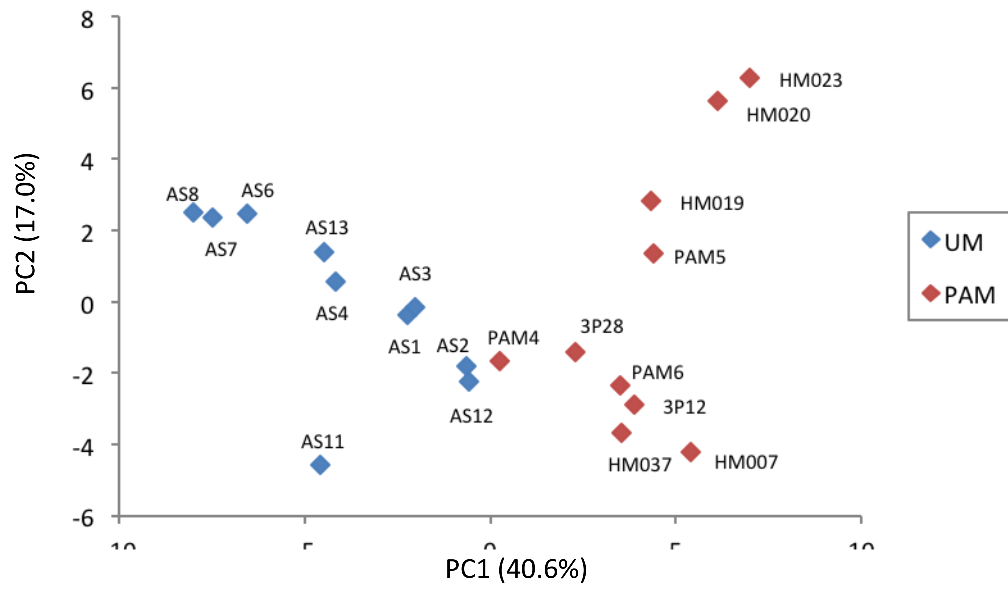


Figure 4:

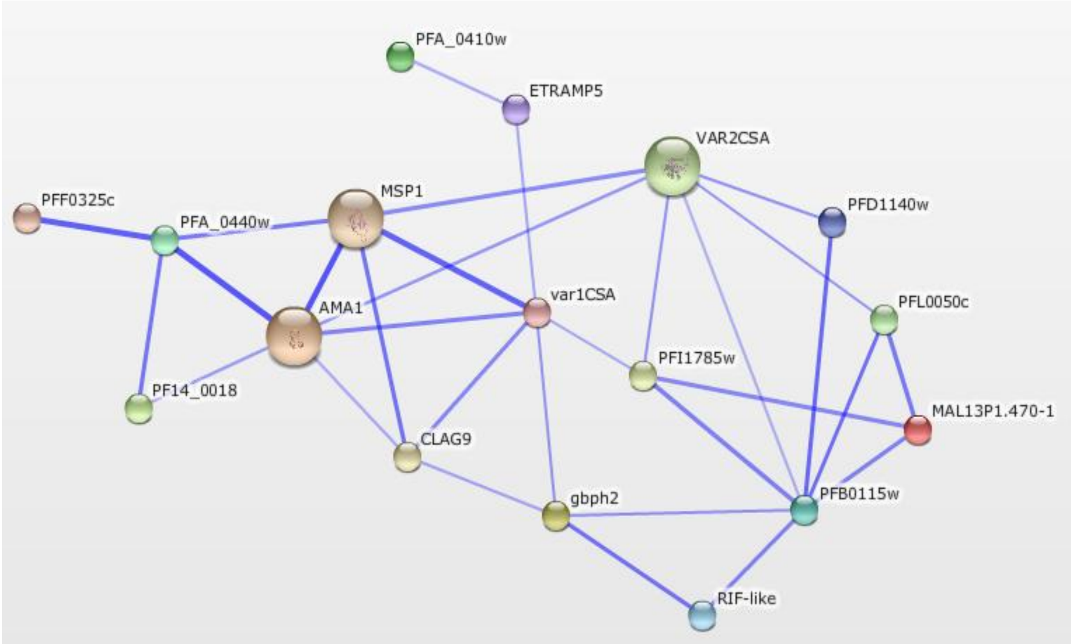


Table 1: List of *Plasmodium falciparum* proteins expressed and localized at the membrane of erythrocyte infected by mature stages that have been identified by LC-MS/MS. The expression of selected erythrocyte membrane skeleton proteins is shown as control.

Group	Sample	Description	CLAG 3.1	CLAG 3.2	CLAG 9	PfEMP3	KAHRP	MAHRP2	RhopH2	RhopH3	skeleton binding protein 1	α spectrin, alpha [Homo sapiens]	spectrin, beta [Homo sapiens]
Uncomplicated malaria (UM)	AS1	Score	82	82		239	1189		157	34	64	7385	4140
		Peptides	4	4		18	88		9	3	5	414	285
	AS2	Score			58	100	2056		58	46	41	3991	3022
		Peptides			2	6	94		9	2	2	226	193
	AS3	Score	119	264	96	119	564	38	1091	336	106	6306	4331
		Peptides	10	19	9	7	25	3	71	29	5	337	216
	AS11	Score	125	228	139	493	1318	97	943	924	177	7945	7049
		Peptides	9	13	12	38	65	5	52	40	19	376	276
	AS12	Score				138	575			114	99	9612	7549
		Peptides				7	27			5	9	454	310
	AS13	Score	50	50		763	2449	59	386	82	90	8619	3956
		Peptides	8	8		56	144	3	30	10	3	452	205
	AS4	Score	71	71		154	488		300	371	44	10717	8235
		Peptides	5	5		10	28		20	18	5	478	324
	AS6	Score	76	188	74	800	617	52	1424	1351	115	22976	15762
		Peptides	4	8	5	62	52	3	75	69	10	1023	703
	AS7	Score	92	274	123	671	1225	86	1225	566	77	9016	6993
		Peptides	6	11	8	46	59	5	73	33	5	449	284
	AS8	Score	47	89	25	655	1470	76	941	640	184	9260	6931
		Peptides	5	7	4	40	75	4	59	37	16	436	317
Placental blood (PAM)	3P12	Score	43	43		40	42		1251	53		1546	2574
		Peptides	3	3		2	2		56	2		149	109
	3P28	Score			38		113	108		63	70	443	572
		Peptides			4		10	3		5	8	41	29
	HM019	Score	378	536	228	334	453	147	1785	714	279	11372	8789
		Peptides	18	23	13	15	11	4	76	31	5	380	304
	HM037	Score							63	151		1505	859
		Peptides							4	3		68	36
	HM007	Score	242	209	138	300	379	244	752	709	199	7249	6127
		Peptides	12	14	16	20	22	10	57	32	16	264	253
	HM020	Score	760	1010	38	566	2246	151	1587	1086	54	26699	27857
		Peptides	44	54	4	37	86	7	83	54	6	1220	1138
	HM023	Score	963	1259	54	993	3135	262	1913	1166	84	26785	28275
		Peptides	55	72	9	48	123	11	101	63	12	1162	1108
	PAM4	Score	179	224	168	90	1785		491	211	55	3455	2911
		Peptides	12	13	11	7	85		40	9	4	195	171
	PAM5	Score	349	421	167		291	73	940	487	74	5135	3762
		Peptides	22	29	15		15	4	62	32	8	233	171
	PAM6	Score	65	67			374		216	30		8738	6685
		Peptides	3	4			16		15	2		497	401

Table 2: Results of filter-based feature selection methods (a) and supervised classification analysis on the different sets of proteins before and after filtering (b).

a)

Filter method	No. of features selected	Unique to one filter	Common to two filters	Common to three filters
Fisher's ANOVA	61	9 (15%)	36 (59%)	16 (26%)
Runs test	24	6 (25%)	2 (8%)	16 (67%)
ReliefF	72	20 (28%)	36 (50%)	16 (22%)

b)

	Classification accuracy (%) before filtering	Classification accuracy (%) after filtering			
		Fisher's ANOVA	Runs	ReliefF	53-protein set
No. of features	454	61	24	72	53
Classification algorithm					
Classification tree algorithm (CART)	75%	**90%**	85%	**90%**	**90%**
k-nearest neighbors	75%	85%	80%	85%	**95%**
Multilayer perceptron neural network	70%	75%	80%	65%	**95%**
Radial basis function neural network	75%	75%	80%	80%	**85%**
Support vector machine	80%	85%	85%	85%	**95%**

Classification accuracy was estimated as the overall number of correctly classified samples divided by the total number of samples through a leave-one-out cross-validation procedure. The highest classification accuracy achieved by each of the five classification algorithms is shown in bold

Table 3: Description of proteins discriminating uncomplicated malaria (UM) and pregnancy-associated malaria (PAM) samples

Accession	Description	MW kDa	ΣCoverage	Σ# Peptides	Molecular function	Cellular component	Clinical group	Group	PEXEL motif	No.TM domains	*P* values
23496536	RIFIN	40.5	13.82	5	catalytic activity; metal ion binding; transporter activity	extracellular; membrane	UM	A	Yes	3	0.016
86170346		41.3	7.28	4							0.007
124505671		41.4	12.47	8							0.006
23497064		39.7	12.86	5							0.002
124803060		41.7	6.18	3							0.002
59803938	MSP1	19.9	58.19	11	-	membrane	PAM	-	No	0	0.026
23496537	PFL0030c	355.0	7.82	22	protein binding; receptor activity; signal transducer activity	extracellular; membrane	PAM	UPS E	Yes	1	0.041
VARDOM	HB3var2csaA	336.3	6.54	17							0.010
154359874	VAR2CSA	284.1	8.55	20							0.010
154359884	VAR2CSA	314.4	6.82	17							0.004
164419782 164419780	PFI1785w	43.2	34.99	9	unknown	membrane	PAM	PHISTb	Yes	1	0.004
258597612	PF14-0018	55.8	30.54	13					-		0.019
124505801	PFA-0410w	254.9	26.88	42	unknown	-	PAM	-	No	0	0.032
296004626	PFF0325c	280.2	4.64	11						2	0.022
124800702	PFB0115w	141.8	22.90	27	catalytic activity; transporter activity	membrane				1	0.004

MW: Molecular Weight; ΣCoverage: percentage of the protein sequence covered by all identified peptides from all samples; Σ# Peptides: Number of different detected peptides in all samples and matching with the protein sequence; No. TM domains: number of transmembrane

domains; *P* values: comparison of abundance of each proteins between two groups by Wilcoxon rank sum tests. *P* values <0.05 were considered as significant.