Probabilistic data integration identifies reliable gametocyte-specific proteins and transcripts in malaria parasites

Lisette Meerstein-Kessel[1,2], Robin van der Lee[1#a], Will Stone[2,3], Kjerstin Lanke[2], David A Baker[4], Pietro Alano[5], Francesco Silvestrini[5], Chris J Janse[6], Shahid M Khan[6], Marga van de Vegte-Bolmer[2], Wouter Graumans[2], Rianne Siebelink-Stoter[2], Taco WA Kooij[2], Matthias Marti[7], Chris Drakeley[3], Joseph J. Campo[8], Teunis JP van Dam[1#b], Robert Sauerwein[2], Teun Bousema[2¶], Martijn A Huynen[1¶*]

[1] Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud university medical center, Nijmegen, The Netherlands
[2] Department of Medical Microbiology, Radboud university medical center, Nijmegen, The Netherlands
[3] Department of Immunology and Infection, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom
[4] Department of Pathogen Molecular Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom
[5] Dipartimento Malattie Infettive, Istituto Superiore di Sanità, Rome, Italy
[6] Department of Parasitology, Leiden University Medical Center, Leiden, The Netherlands
[7] Wellcome Trust Center for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, College of Medical Veterinary & Life Sciences, University of Glasgow, Glasgow, Scotland, United Kingdom
[8] Antigen Discovery Inc., Irvine, California, USA
[#a] Current Address: Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, BC Children's Hospital Research Institute, University of British Columbia, Vancouver, BC, Canada
[#b] Current Address: Theoretical Biology and Bioinformatics, Department of Biology, Utrecht University, Utrecht, The Netherlands

* Corresponding author
Email Martijn.Huijnen@radboudumc.nl (MAH)

[¶] These authors contributed equally to this work.

Abstract

Plasmodium gametocytes are the sexual forms of the malaria parasite essential for transmission to mosquitoes. To better understand how gametocytes differ from asexual blood-stage parasites, we performed a systematic analysis of available 'omics data for *P. falciparum* and other *Plasmodium* species. 18 transcriptomic and proteomic data sets were evaluated for the presence of curated "gold standards" of 41 gametocyte-specific versus 46 non-gametocyte genes and integrated using Bayesian probabilities, resulting in gametocyte-specificity scores for all *P. falciparum* genes.

To illustrate the utility of the gametocyte score, we explored newly predicted gametocyte-specific genes as potential biomarkers of gametocyte carriage and exposure. We analyzed the humoral immune response in field samples against 30 novel gametocyte-specific antigens and found five antigens to be differentially recognized by gametocyte carriers as compared to malaria-infected individuals without detectable gametocytes. We also validated the gametocyte-specificity of 15 identified gametocyte transcripts on culture material and samples from naturally infected individuals, resulting in eight transcripts that were >1000-fold higher expressed in gametocytes compared to asexual parasites and whose transcript abundance allowed gametocyte detection in naturally infected individuals.

Our integrated genome-wide gametocyte-specificity scores provide a comprehensive resource to identify targets and monitor *P. falciparum* gametocytemia.

Introduction

Despite a decrease in malaria incidence and mortality over the past two decades, malaria remains a major global health challenge [1,2]. Furthermore, the emergence and spread of insecticide resistance in mosquitoes [3] and artemisinin resistance in *Plasmodium falciparum* (*Pf*) [4–6] threaten recent gains in malaria control. The decline in malaria burden and the necessity to contain artemisinin-resistance have increased interest in malaria elimination that may require interventions that specifically aim to prevent malaria transmission. Malaria transmission depends on male and female gametocytes, the sexually reproducing forms of the *Plasmodium* parasite that are ingested by blood-feeding *Anopheles* mosquitoes. In the mosquito gut, gametocytes may complete the parasite's reproductive cycle and, following sporogonic development, render the mosquito infectious. Factors that govern gametocyte production and infectivity remain poorly understood. Whilst recent studies have shed light on the processes controlling gametocyte commitment [7,8], commitment and maturation of gametocytes may differ between infections and over the course of infections, under influence of environmental and host factors [9,10]. A better understanding of gametocyte dynamics during infections, as well as the development of tools to monitor or target gametocytes, may be informed by high-throughput protein and transcriptome studies [11,12]. In the past 15 years, a number of large-scale studies on *Plasmodium* gametocytes have been reported: the proteome of *Pf* and the rodent malaria parasite *Plasmodium berghei (Pb)* have been examined by mass spectrometry [13–21], and the transcriptome of both species by microarray and RNA sequencing [15,20,22–28]. These studies differed in their focus and resolution in examining (sexual) developmental stages and each faced challenges in detecting low abundance proteins [29] and by the purity of parasite populations [13,14]. The use of fluorescent parasites and fluorescence-assisted sorting of staged parasites have recently permitted a better discrimination of proteins in either male or female gametocytes [16,17,20] and have allowed more detailed comparisons of *Plasmodium* life-stages. However, individual studies are still vulnerable to imperfect sample purity, and other sources of uncertainty such as correct gene identification for accurate peptide assignment. These technical and methodological challenges lead to discrepancies between individual studies and hamper firm conclusions about gametocyte-specificity of proteins and transcripts.

We utilized the numerous published proteomics and transcriptomics *Plasmodium* data sets in a comprehensive data integration framework to obtain a consensus of gametocyte-specific transcripts and proteins. Our data integration approach is an adaptation of the naïve

Bayesian classifiers that have previously been applied in the prediction of protein interactions and components of cellular systems [30,31]. The framework calculates probabilities that any given transcript or protein is gametocyte-specific given the evidence presented across the total of transcriptomics and proteomics data. A key aspect of the methodology is that it takes into account the predictive power of each contributing data set: it assigns weights to data sets based on their ability to distinguish gold standard lists of gametocyte and asexual proteins. These we have constructed using existing literature where life-stage specificity was confirmed using classical "non-omics" approaches (e.g. protein detection in immunofluorescence-assays, functional/genetic studies), followed by expert curation. The most informative data (from datasets with the highest discriminative power against the gold-standard lists) will thus contribute most to the predictions, while less informative data are down-weighted. This allows for (i) the resolution of conflicting evidence without disregarding data, and (ii) the construction of a transparent scoring system in which the relative contribution of each data set is directly visible.

Using this approach, we propose a robust gametocyte-specificity score for all *Pf* genes that allows a consensus list of gametocyte-specific genes at protein and transcript level. We illustrate the utility of our findings by examining naturally acquired responses to newly identified gametocyte-specific proteins in gametocyte-carriers and non-carriers by protein microarray. In addition, we confirmed gametocyte-specificity for a selection of gametocyte-specific transcripts using culture material from geographically distinct *Pf* strains and samples from naturally infected malaria patients.

Results

Weighted integration of proteomics and gene expression data

Using Bayesian statistics, we integrated *Plasmodium* mass spectrometry and transcript datasets from 18 different studies (Table 1) on *P. falciparum (Pf;* n=14*)*, *P. berghei (Pb;* n=3*)* and *P. vivax (Pv*; n=1*)*. Since gametocyte biology differs between *Plasmodium* species, scores were calculated for the total set of *Plasmodium* studies and for *Pf* only. Unsupervised clustering of genes based on peptide counts or mRNA expression resulted in grouping according to data acquisition method rather than parasite stage (Fig 1), illustrating the necessity of a supervised approach to discriminate between gametocyte-specific and non-gametocyte genes. To objectively assess the value of individual data sets and allow their assembly into a gametocyte-specificity score, we created a gold standard that served as a

benchmark for every sample. This gold standard was collected from literature review and comprises two lists; one of asexually expressed proteins, mainly blood stage but also sporozoite and liver stage, and one of known gametocyte proteins (Supplementary Table S1). A gametocyte-specificity score was then derived for each gene by comparing its expression in all studies to the relative expression of the gametocyte and asexual gold standards in those samples (Supplementary Fig S1-2). Proteins or RNAs detected in a study with high discriminative power for gametocyte and asexual gold standard genes (ratio of gametocyte to asexual gold standard genes) received higher gametocyte-specificity scores than those detected in a study with lower discriminative power. The individual log-transformed scores per gene were combined for proteomics and transcriptomics data separately. Scores for *Pf*-only studies (Supplementary Table S2) and all combined data sets were highly correlated (Fig 2D; Pearson's r=0.9867 and r=0.9514 for proteomics and transcriptomics, respectively) and the latter were used in the remainder of the manuscript. The distribution of scores for proteins and transcripts are presented for the two sets of genes of the gold standard as well as for all other genes (Fig 2A). As expected, the gametocyte and asexual gold standard set of genes are perfectly separated by their respective proteomics-derived scores and show only little overlap in their transcriptomics-derived scores (Fig 2B). The shift in the density peak of the proteomics compared to transcriptomics is due to an inherent property of the method that gives a negative score for gametocyte-specificity to all proteins that were not detected in the proteomics studies (n=1583). The highest scoring 100 genes for proteomics and transcriptomics contained 26 (63.4 %) and 15 (36.6 %) of the 41 gold standard gametocyte genes, respectively (Fig 2C), indicating both the respective discriminating power of the gold standard and that many other genes are as specific for gametocytes as the highest gold standard representatives.

Translationally repressed genes are common in late stage female gametocytes [32,33] and are detectable by high transcriptomics and low proteomics score. In our analysis 461 genes have this profile (Supplementary Table S3), including genes that are known to be translationally repressed like Pfs28 [32,34] and 186 genes with a previously reported bias towards expression in female gametocytes [20].

Cross-validation illustrates the improved predictive power of the integrated data. Ten-fold cross-validation was performed using random subsamples of the gold standard lists to predict the ranks of left-out genes. The resulting proteomics ranking shows near perfect sensitivity, with all but two gold standard gametocyte genes ranking higher than the gold standard asexual genes (Fig 3). The added value of our integrated approach is illustrated by the receiver operating characteristic curve where the integration of data sets gave higher sensitivity and areas under the curve for both proteomics and transcriptomics than any

individual study (Fig 3). Using the Bayesian integration based on the complete gold standards, we ranked all *Pf* proteins by giving them a gametocyte-specificity score (Supplementary Table S2). All proteins with a score >5 (n=602) were considered gametocyte-specific. Most of these have not consistently been described as "specific" or "enriched" in gametocytes in the original data sets (Fig 4A and Supplementary Table S4). Previous studies defined 315 [13] to 1725 [20] proteins as gametocyte-specific for *Pf*. Not only did our integrated approach lead to a better recovery of gold standard listed known gametocyte proteins, we also identified 178 genes with undescribed function as gametocyte-specific (Supplementary Table S2). We further identify a number of proteins as gametocyte-specific even though they had been reported as asexual by previous studies (Fig 4B). A recent proteomics study of male and female *Pf* gametocytes [35], not included in our original analysis, was used to test the robustness of our scores. When we included this data set in our final Bayesian proteomics scores, both gametocyte scores and gene ranks before and after addition of this data set were highly correlated (Pearson's r=0.997 and Spearman's rho=0.995, respectively). Furthermore, the top 100 gametocyte proteins did not change and the top 602 proteins were 96% identical (578 of 602). Taken together, cross-validation and independent data suggest that the integrated gametocyte-specificity score is robust and contains potential novel gametocyte markers.

Predicted gametocyte-specific proteins are recognized by gametocyte-carriers.

As an illustration of the utility of gametocyte-specific proteins as markers of gametocyte exposure, we utilized protein microarray data from a study that aimed to characterize the immune profile associated with transmission-reducing immunity in naturally infected gametocyte carriers (Stone, Campo *et al.* 2017 accepted manuscript [36]). For the current study, we compared responses to our gold standard gametocyte genes (n=40) and novel gametocyte genes from our 100 highest scoring proteins that were on the array (n=30). Antibody prevalences for these genes were compared between Gambian gametocyte carriers and Gambians who carried asexual parasites but not gametocytes as determined by microscopy. Antibody responses to the predicted gametocyte-specific proteins were significantly higher in gametocyte carriers (p=0.005), while for the gold standard antigens this difference was less significant (p=0.058) (Fig 5A, Mann-Whitney U test). When antigens were analysed individually, a significantly higher antibody prevalence in gametocyte carriers was detected for five novel gametocyte antigens (Fig 5B, p<0.05 in Fisher's exact, corrected for multiple testing, Supplementary Table S5, Supplementary Fig S3). Only two of these five have an assigned function – a DNA ligase, and Gamete egress and sporozoite-traversal protein (GEST). For two of the three remaining *Plasmodium* proteins, we were able to predict a function based on homology, using the sensitive homology detection tool HHpred [37].

PF3D7_1251000 is homologous to the co-chaperone HSP20 heat shock protein and PF3D7_1439600 is homologous to the MLRQ subunit of complex IV of the oxidative phosphorylation, underlining the enrichment in mitochondrial proteins as discussed below (Supplementary Table S6 includes homology predictions for all conserved, highly gametocyte-specific *Plasmodium* proteins).

Gametocyte-specific RNA transcripts detect (submicroscopic) gametocyte carriage

Of the 100 highest-scoring transcripts, 15 non-gold standard candidates were selected for qRT-PCR validation based on their gametocyte scores in a preliminary analysis (Table 2). Mature gametocytes of four *Pf* strains from different geographical origins were compared to asexual blood stage parasites. The minimum transcript abundance difference ($Ct_{Asexuals}$ - $Ct_{Gametocytes}$) ranged from 4.76 to 14.95 (Fig 6A and Supplementary Table S7 qPCR & primers), reflecting 27.1 to 31,500-fold higher transcript numbers in gametocytes compared to asexual parasites and confirming pronounced upregulation of all selected targets in gametocytes. With a very conservative threshold of 1,000-fold enrichment in three of the four strains tested (Fig 6A), eight of the 15 tested transcripts were highly specific to gametocytes. Transcript abundance in ring-stage parasites was assessed and compared to Pfs25 mRNA, an established and highly abundant yet intron-less female gametocyte specific transcript [11,38]. Five out of eight gametocyte specific transcripts were undetectable in asexual ring stages at $\leq 10^5$ parasites/mL, similar in specificity to Pfs25 (Fig 6B); the five most sensitive gametocyte markers detected gametocytes across the range of 10^2 – 10^6 gametocytes/mL (Fig 6C). In RNA samples from a previously reported clinical trial conducted in Kenya [39], all eight gametocyte markers detected gametocytes at densities below 10^3/mL (Fig 6D).

Gametocyte-specific proteins are enriched for cytoskeletal movement and metabolism functions

To uncover novel characteristics underlying gametocyte function, we analyzed over-represented gene ontology (GO) terms in our integrated consensus gametocyte proteins. The 100 highest ranked proteins were examined for enrichment of GO terms that reflect specific biological functions. Microtubule based movement, metabolism of carboxylic acids and metabolism of nucleic acids were highly enriched among gametocyte proteins (Supplementary Fig S4). Four out of the six putative *Pf* dynein heavy chain proteins are found back among the 100 most gametocyte-specific proteins, alongside a tubulin gamma chain and a tubulin chaperone. The importance of DNA elongation and ligation processes is reflected in GO term associations as well as in antibody response to a DNA ligase (Fig 5B). The “classic” GO term enrichment calculation was complemented by a rank-based gene set

enrichment analysis (GSEA). GSEA uses all *Pf* genes and their respective (proteomics derived) gametocyte-specificity score and thus does not include an arbitrary cutoff of the proteins that are or are not gametocyte-specific. It confirmed the above-mentioned results, and in addition to those the terms “mitochondrial protein complex“ (GO:0098798) and “TCA” (GO:0006099) were enriched, stressing both mitochondrial location and processes. Although male gametocytes carry pre-synthesized proteins to rapidly form eight motile gametes upon activation in the mosquito midgut, we did not observe flagellum associated terms in the GSEA. The reason for this is that current GO term annotation for *Pf* has only one gene with the “cilium” GO term for cellular component (GO:0005929; PF3D7_1025500), one for “axoneme” (GO:0005930; PF3D7_0828700) and none with the biological-process terms “cilium or flagellum-dependent cell motility” (GO:0001539) or “axoneme assembly” (GO:0035082). We supplemented the GO annotation with a list of 28 *Pf* cilium genes (Methods). The newly assembled “cilium” GO term now acquired the highest enrichment score in the GSEA (Supplementary Fig S5). This may reflect the formation of the flagella of the microgamete but may also (partially) reflect intracellular trafficking and or be associated with genome replication as the term has overlap with the genes annotated for microtubule processes (Supplementary Table S5 GO terms).

Discussion

Combining proteomic and transcriptomic data from 18 sources, we present an integrated consensus score for gametocyte-specific proteins and transcripts. We predict 602 gametocyte-enriched proteins of which 186 are currently without ascribed function. We illustrate the potential utility of our gametocyte score by providing evidence for differential recognition of gametocyte proteins by naturally infected gametocyte carriers and the sensitive detection of mRNA of novel gametocyte transcripts in field samples.

The gametocyte proteome of *P. falciparum (Pf)* has been assessed repeatedly. Individual lists of gametocyte-specific proteins [13,14,16,17,19,20] have unavoidable limitations related to comparator (asexual) parasite stages, sample purity, assay sensitivity and arbitrary cut-offs used to define gametocyte-specificity and show only partial agreement. To acquire a more robust gametocyte-specificity score, we integrated data from these individual studies, along with studies of purified asexual parasites and related *Plasmodium* species. Including gene expression data from multiple species generally increases the likelihood that the combined gene expression data reflect underlying biology, as observed in the *Apicomplexa* [40,41]. We applied a Bayesian classifier first applied to ‘omics data by Jansen and colleagues [30] and

adapted by Van der Lee and colleagues to identify genes involved in anti-viral immune responses [31]. The probabilistic approach combines the evidence from all studies in an unbiased way, without giving *a priori* preference of one study over another. Instead, the measurements of all studies were weighted inherently during the scoring process by assessing the retrieval of a gold standard set of genes. As these gametocyte and asexual gold standard sets are of central importance to the study, they have undergone expert curation (see Methods and Acknowledgments). The power of this integrative approach lies not only in weighting data sets by the retrieval of gold standard genes but also in the opportunity to exclude proteins from the gametocyte-specific list by appreciating their presence in (other) asexual samples. A further strength of the approach is that it allows the ranking of gametocyte proteins that have only been reported in a subset of studies. Our integration of data sets reveals that 602 proteins are likely to be specific to gametocytes although very few gametocyte-specific proteins were detected in every underlying dataset and seven proteins had never before been reported as gametocyte-specific.

A general limitation of all mass spectrometry (MS) studies is their bias toward highly abundant proteins. Proteins with low-level expression may be missed in a bulk proteome analysis. After integration of the MS studies listed in Table 1, 1583 *Pf* proteins were never detected, representing approximately 28% of all proteins encoded by *Pf*. Some of these might be of too low abundance or expressed during sporozoite or gamete, ookinete and liver stage, which are underrepresented or not included in our data, respectively. New advances in MS that include the sensitive detection of peptides from currently understudied *Plasmodium* life stages may shed light on these currently uncharacterized genes. In addition, approaches that focus specifically on post translational modifications like phosphorylation of proteins as has been done for asexual parasites [42–45] may add new lines of evidence towards gametocyte-specific functions of proteins. Our approach suggests that the currently available MS data is sufficiently comprehensive to identify stage-specific proteins when analysed in an integrative approach. We examined this directly by incorporating a new *Pf* gametocyte MS study [35] in our scoring. The authors reported 44 new gametocyte-specific proteins that were not reported by earlier studies. We compared this data set to our integrated data set and found 24 of the 44 had been detected in one or more erythrocytic stages or sporozoites [14,18,42,46,47] while 11 others had been identified in a (single) gametocyte sample before (Supplementary Fig S6). Importantly, the scores and top 100 gametocyte genes remained unaltered by integrating this new dataset.

The ranking of gametocyte-specificity that we provide here can i) aid in understanding the biology of this life stage and ii) improve diagnostics related to gametocyte exposure and carriage. Regarding gametocyte biology, our high-ranking gametocyte-specific genes are

enriched for mitochondrial, metabolism and microtubule processes and DNA replication, supporting the quality of the data integration. The enrichment of mitochondrial localization and process is consistent with what we know about the enlarged mitochondrion of gametocyte stages [48] and increased activity of the citric acid cycle [49]. DNA replication terms are highly enriched which is consistent with what happens in the subsequent life stage in which the (micro)gamete rapidly duplicates its genomic DNA three times. Regarding the use of the gametocyte score to inform gametocyte diagnostics, diagnostics can directly detect nucleic acids specific to gametocytes [11] or detect antibody responses reflecting past/recent exposure as is increasingly used for asexual *P. falciparum* and *P. vivax* parasites [50,51]. We use our integrated gametocyte list to explore its utility for both approaches. We validated 15 transcript targets in four different *Pf* strains, comparing transcript abundance in gametocytes and asexual parasites. All tested targets were enriched in gametocytes. Five targets were tested for their sensitivity and can recognize 100 gametocytes/mL, while the signal is undetectable when fewer than 10^5-10^6 ring-stage parasites/mL are present. In practical terms, these markers may be used to reliably detect gametocytes at densities well below the microscopic threshold of detection in samples without high-densities of asexual parasites, similar to the gametocyte marker that is currently most widely used, the female gametocyte-specific Pfs25 [38].

As an alternative approach to the detection of gametocyte carriage in populations, we utilized a gametocyte-enriched protein microarray (Stone, Campo *et al*. 2017 accepted manuscript [36]) to determine antibody responses to genes that we here describe as highly gametocyte-specific. The bacterial expression system used for the array has known limitations with the expression of conformational proteins [52] and should thus be considered a 'rule in' rather than 'rule out' approach to immune recognition. Moreover, the array was constructed with the aim of detecting surface proteins or exported proteins whilst our list does not require these characteristics. Only 30 of our top 100 novel gametocyte antigens were thus printed on this array. Antibody responses to five gametocyte proteins were significantly more prevalent in gametocyte carriers than in carriers of the asexual blood stage only. This is the first evidence that antibody responses may be indicative of current gametocyte carriage. Importantly, the dichotomization of gametocyte-exposed and non-exposed individuals was based on a single time-point screening for gametocytes by microscopy. Microscopy has a low sensitivity for detecting gametocytes that commonly circulate at low densities [53] and several of the asexual parasite carriers are likely to have had preceding or concurrent low densities of circulating gametocytes. Antibody prevalence in the group classified as gametocyte-negative by microscopy may thus be associated with concurrent low-density gametoctytemia and/or long-lived antibody responses acquired following previous gametocyte exposure. The presently analysed samples thus do not allow

any conclusions on a possible role of submicroscopic gametocyte densities in boosting or maintaining antibody responses to gametocyte antigens. Refined studies with longitudinal sampling and gametocyte detection by sensitive qRT-PCR methodologies are needed to formally assess antibody kinetics in relation to gametocyte exposure and determine whether recent markers of exposure to blood stage antigens [50] can be complemented by a set of markers for recent or long-term gametocyte exposure.

We described the assembly of a curated gold standard set of gametocyte and asexual proteins and used this new resource to rank the likelihood of all *Pf* proteins and transcripts being specific to the gametocyte stage. Data from 18 publicly available studies were integrated to resolve partially conflicting evidence. The resulting consensus lists can be used for guidance of future investigations as we have shown the value of our predictions by in vitro validation.

Materials and Methods

Assembly of a gold standard for gametocyte and asexual proteins to weigh whole proteome/transcriptome data sets

To build a gold standard against which the performance of individual data sets could be assessed, we identified proteins that are known to be expressed in either asexual parasites (mostly blood stage, also including sporozoites and liver stage) or gametocytes. This list was initially informed by literature review (Supplementary table S1) for expression in the respective stages as detected by immunofluorescence assays and/or western blot, supplemented with *P. falciparum* blood stage or transmission blocking vaccine candidates. This initial list was then communicated with experts (including the authors DAB, PA, FS, CJJ, SMK, TWAK, MM, CD, RS and TB) and edited. If additional proteins were suggested for inclusion in the list, published evidence was requested and examined prior to inclusion of the proteinThe final asexual gold standard list contains 46 proteins; the final gametocyte list contains 41 proteins. These gold standard lists (Supplementary Table 1) represent the balance between very strict inclusion criteria and sufficient set size to evaluate the quality of all data sets integrated. We tested for the detection of these proteins or transcripts in the respective samples, using a Bayesian statistics approach that we have successfully applied previously for genes involved in anti-viral immune defense [31].

Data selection and integration

Data sets that measured protein and transcript abundance in *Pf* gametocytes were balanced with data from other life stages and supplemented with studies of the rodent malaria parasite *P. berghei*. One MS study on *P. vivax* was included as it is based on of *ex vivo* blood material as opposed to all other studies that used *in vitro* cultivated parasites. Unique peptide counts were retrieved from plasmoDB (version 28) in which sequenced peptides from the published studies are always mapped to the most recent genome annotation, or supplementary material of the respective studies. Many aspects determine how well a protein is represented in proteomics data that are obtained via MS, like its length or posttranslational modifications. Remapping original MS data to newly annotated genomes improves the quality of the predicted proteins [19]. We were however not able to retrieve those data from the studies [13,14,16] and therefore decided to take those proteins at face value. Notice that also these early studies contribute significantly to our integrated lists.

Expression percentiles were retrieved from plasmoDB (version 28) or calculated from raw data in the respective supplementary material. Gametocyte samples were summarized if applicable (using the maximum peptide count/expression percentile of different stages or male and female gametocytes) as were asexual samples, only considering the highest expression in any sample or time point.

For MS and transcriptomics data sets, separate scores for gametocyte-specificity of any *Pf* gene have been calculated. In brief, protein or transcript expression has been categorized from absent to high expression levels as given by number of unique peptides or expression percentiles, respectively. For each of the respective bins, a score was calculated depending on the relative retrieval of gametocyte and asexual gold standard genes. The log ratio of these retrieved genes defined the score for all other genes within the same bin. The final gametocyte score calculates as the prior probability of a gene being gametocyte-specific that is updated using the contributions of the data sets:

$$GametocyteScore = log_2\left(\frac{P_{gct}}{P_{\sim gct}}\right) + \sum_{i=1}^{n} log_2\left(\frac{P(data_i|Gct)}{P(data_i|\sim Gct)}\right) \qquad (1)$$

$$with\ \frac{P(data_i|Gct)}{P(data_i|\sim Gct)} = \frac{gametocyteGS_i}{asexualGS_i} \qquad (2)$$

where gametocyteGS and asexualGS are the fractions of retrieved gametocyte and asexual gold standard genes in sample i, respectively. We used a pseudocount of 1 if necessary to prevent division by zero if none of the gold standard genes was retrieved in this specific sample and bin. It was assumed that the likelihood of a gene to be either gametocyte or asexual specific is equally high, thus the (log-transformed) prior equals 0 and the final score

depends solely on the integrated data. In the selection of a set of proteins that we assigned to be gametocyte-specific we chose a cutoff score of 5.0 (proteomics-derived). The cutoff score of 5.0 can be interpreted as: a gene has to be $2^5 = 32$ times more likely to be gametocyte-specific than asexual specific. The score of 5.0 was based on the behavior of the gold standard genes. Out of the 41 gametocyte gold standard genes, 37 have a score higher than 5.0, while none of the asexual gold standard genes do.

When applicable, genes from *Pb* and *Pv* were treated as their respective *Pf* orthologs as retrieved from plasmoDB [54], to be able to integrate all data sets. When no ortholog is known, the respective non-*Pf* data sets did not contribute to the score of this particular gene. Scores using *Pf* data exclusively were also calculated (Supplementary Table S2 includes all scores and rankings with expression information from all integrated studies).

Cross-validation of the scoring method

We performed a ten-fold cross-validation to assess the predictive performance of the integrated gametocyte-specificity score (i.e. its ability to discriminate known gametocyte vs. asexual genes). For that, we subsampled both gold standard gene sets ten times (folds), without replacement (i.e. each gene is selected exactly once). Then for each fold we re-weighed and integrated the data sets based on nine-tenth of gold standard genes, and collected the ranks of the one-tenth of genes that were left out in that particular fold. A ROC curve was constructed based on those ranks. Using the same strategy, ROC curves for individual data sets that comprised both gametocyte and asexual samples were constructed for comparison.

Protein microarray to measure humoral immune responses

A protein microarray that was enriched for gametocyte proteins was produced and probed as described earlier for a study aiming to unravel the immune signature of naturally acquired transmission-reducing immune responses in gametocyte carriers (Stone, Campo *et al*. 2017 accepted manuscript [36]). As a control group, Gambian asexual parasite carriers without gametocytes detectable by microscopy were included in the probing. For the current study array data from this control group (n=63) and Gambian gametocyte carriers were used (n=164) [55–60]. All of these 227 individuals were sampled during a period of intense malaria transmission intensity in The Gambia and likely had (multiple) previous malaria infections [55,61]. For these populations, responses to 30 newly defined highly gametocyte-specific antigens (from 24 genes) were compared between gametocyte carriers and non-carriers (Mann-Whitney U test). Seropositivity for each of the antigens was determined using a mixture model-based cutoff and related to gametocyte carriage using Fisher's exact test,

corrected for multiple testing (Benjamini-Hochberg) of a total of 70 antigens (including 40 antigens from the gametocyte gold standard).

Transcript abundance in different life stages and strains

The abundance of 15 predicted gametocyte-specific targets was measured in asexual parasites and gametocytes of four different *Pf* strains from *in vitro* culture. The targets were selected from the 100 highest scoring transcripts to account for uncertainties about the absolute scoring of transcriptomics data with a protein-based gold standard. We do not assume a clear hierarchy between these top 100 scoring transcripts and consider any of these genes highly gametocyte-specific. The 15 highest-ranking non-gold standard genes were selected based on a preliminary analysis of the data, and contain genes that are currently not annotated as well as genes with known protein function in gametocytes (PUF1 [62] and Ccp4 [63]) . In the final generation of the gametocyte-scores, all validation genes were retained in the top 100 scoring genes. The *Pf* strains used are of West African (NF54, NF166, NF175) and Southeast Asian origin (NF135). All strains were cultured and synchronized as described previously [20]. Using established standard curves, the same concentrations of parasites were compared for Ct values in qRT-PCR (for primers, see Supplementary Table S7). Extracted nucleic acids were DNase-treated before reverse transcription when introns were absent from the targets. Initial comparison was between mixed asexual blood stage parasites (considering the lowest Ct measured in any strain and replicate) and stage V gametocyte (highest Ct measured per strain and replicate). Promising targets with a high $Ct_{Asexuals}$ - $Ct_{Gametocytes}$ were further examined in serial dilutions of stage V gametocytes and synchronized asexual material of the strain NF54 (10, 20, 30, 40 hours post invasion, resembling early rings, late rings, trophozoites and schizonts, respectively). All qRT-PCR reactions were analyzed in technical triplicates, from biological triplicates (NF54) or duplicates (remaining strains).

RNA samples were used from a clinical malaria trial conducted in Western Kenya [39]. Samples from days 3 and 7 after treatment were selected to ensure a range of (low-density) gametocyte carriage to test qRT-PCR sensitivity.

Go term enrichment in top 100 proteins or rank-based enrichment

Current GO term annotation for *Pf* was retrieved from plasmoDB (release 30) and analyzed using the topGO R package [64] for the enrichment of terms in the 100 highest scoring proteins versus all *Pf* proteins. Semantic clustering of the significant GO terms in Biological Process ontology was done with the Revigo webtool for *Pf* [65]. Second, gene set enrichment analysis (GSEA) based on all *Pf* proteins and their ranks and scores was performed using the software available at http://software.broadinstitute.org/gsea/downloads.jsp. Based on the

cilium genes reported by the Syscilia consortium [66], we assembled a "cilium" GO term of mixed ontology (GO:9999999) for *Pf* with 28 predicted orthologs (Supplementary Table S6).

Data availability

All data generated or analysed during this study are included in this published article (and its Supplementary Information files).

- The gold standard lists (Supplementary Table S1)
- Bayesian gametocyte scoring for proteomics and transcriptomics data. Includes *Pf* only-scores and expression values for any gene and individual data set (Supplementary Table S2)
- Potential translationally repressed genes with high transcriptomics score (>7) and low proteomics score (<-10, Supplementary Table S3)
- Overview over previously reported gametocyte-specificity per study (Supplementary Table S4)
- Seroprevalence in The Gambia in gametocyte carriers and non-carriers (Supplementary Table S5)
- Function predictions for highly gametocyte-specific proteins with lacking annotation (Supplementary Table S6)
- Transcript validation in 15 targets, including primer sequences for qRT-PCR (Supplementary Table S7)
- GO term analyses with cilium genes (Supplementary Table S8)

References

1. WHO. Reversing the Incidence of Malaria 2000–2015. in (2015).
2. Bhatt, S. *et al.* The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature* **526,** 207–11 (2015).
3. Ranson, H. & Lissenden, N. Insecticide Resistance in African Anopheles Mosquitoes: A Worsening Situation that Needs Urgent Action to Maintain Malaria Control. *Trends Parasitol.* **32,** 187–196 (2016).
4. Ashley, E. A. *et al.* Spread of Artemisinin Resistance in *Plasmodium falciparum* Malaria. *N. Engl. J. Med.* **371,** 411–423 (2014).
5. Takala-Harrison, S. *et al.* Independent emergence of artemisinin resistance mutations among Plasmodium falciparum in Southeast Asia. *J. Infect. Dis.* **211,** 670–679 (2015).
6. Ménard, D. *et al.* A Worldwide Map of Plasmodium falciparum K13-Propeller Polymorphisms. *N. Engl. J. Med.* **374,** 2453–2464 (2016).
7. Kafsack, B. F. C. *et al.* A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature* **507,** 248–52 (2014).
8. Pelle, K. G. *et al.* Transcriptional profiling defines dynamics of parasite tissue sequestration during malaria infection. *Genome Med.* **7,** 19 (2015).
9. Johnston, G. L., Smith, D. L. & Fidock, D. A. Malaria's Missing Number: Calculating the Human Component of R0 by a Within-Host Mechanistic Model of Plasmodium falciparum Infection and Transmission. *PLoS Comput. Biol.* (2013). doi:10.1371/journal.pcbi.1003025
10. Bousema, T. & Drakeley, C. Epidemiology and infectivity of Plasmodium falciparum and Plasmodium vivax gametocytes in relation to malaria control and elimination. *Clinical Microbiology Reviews* **24,** 377–410 (2011).
11. Joice, R. *et al.* Inferring Developmental Stage Composition from Gene Expression in Human Malaria. *PLoS Comput. Biol.* **9,** 1–13 (2013).
12. Proietti, C. & Doolan, D. L. The case for a rational genome-based vaccine against malaria. *Frontiers in Microbiology* **6,** (2015).
13. Lasonder, E. *et al.* Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. *Nature* **419,** 537–42 (2002).
14. Florens, L. *et al.* A proteomic view of the Plasmodium falciparum life cycle. *Nature* **419,** 520–6 (2002).
15. Hall, N., Karras, M. & Raine, J. D. A Comprehensive Survey of the Plasmodium Life Cycle by Genomic, Transcriptomic, and Proteomic Analyses. *Science (80-.).* **307,** 82–86 (2005).
16. Khan, S. M. *et al.* Proteome Analysis of Separated Male and Female Gametocytes Reveals Novel Sex-Specific Plasmodium Biology. *Cell* **121,** 675–687 (2005).
17. Silvestrini, F. *et al.* Protein Export Marks the Early Phase of Gametocytogenesis of the Human Malaria Parasite Plasmodium falciparum. *Mol. Cell. Proteomics* **9,** 1437–1448 (2010).
18. Oehring, S. C. *et al.* Organellar proteomics reveals hundreds of novel nuclear proteins in the malaria parasite Plasmodium falciparum. *Genome Biol* **13,** R108 (2012).
19. Tao, D. *et al.* Sex-partitioning of the Plasmodium falciparum Stage V Gametocyte Proteome Provides Insight into falciparum-specific Cell Biology. *Mol. Cell. Proteomics* **13,** 2705–2724 (2014).
20. Lasonder, E. *et al.* Integrated transcriptomic and proteomic analyses of P. falciparum gametocytes: molecular insight into sex-specific processes and translational repression. *Nucleic Acids Res.* **44,** gkw536 (2016).
21. Suárez-Cortes, P. *et al.* Comparative proteomics and functional analysis reveal a role of P. falciparum osmiophilic bodies in malaria parasite transmission. *Mol. Cell. Proteomics* (2016). doi:10.1074/mcp.M116.060681
22. Bozdech, Z. *et al.* The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biol.* **1,** 85–100 (2003).

23. Le Roch, K. G. *et al.* Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science (80-.).* **301,** 1503–1508 (2003).
24. Young, J. A. *et al.* The Plasmodium falciparum sexual development transcriptome: A microarray analysis using ontology-based pattern identification. *Mol. Biochem. Parasitol.* **143,** 67–79 (2005).
25. Llinas, M. Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains. *Nucleic Acids Res.* **34,** 1166–1173 (2006).
26. Otto, T. D. *et al.* New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq. *Mol. Microbiol.* **76,** 12–24 (2010).
27. López-Barragán, M. J. *et al.* Directional gene expression and antisense transcripts in sexual and asexual stages of Plasmodium falciparum. *BMC Genomics* **12,** 587 (2011).
28. Otto, T. D. *et al.* A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biol.* **12,** 86 (2014).
29. Wasinger, V. C., Zeng, M. & Yau, Y. Current status and advances in quantitative proteomic mass spectrometry. *Int. J. Proteomics* **2013,** 180605 (2013).
30. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science (80-.).* (2003). doi:10.1126/science.1087361
31. van der Lee, R. *et al.* Integrative Genomics-Based Discovery of Novel Regulators of the Innate Antiviral Response. *PLoS Comput. Biol.* **11,** (2015).
32. Mair, G. R. *et al.* Regulation of Sexual Development of Plasmodium by Translational Repression. *Science (80-.).* **313,** 667–669 (2006).
33. Mair, G. R. *et al.* Universal features of post-transcriptional gene regulation are critical for Plasmodium zygote development. *PLoS Pathog.* **6,** (2010).
34. Miao, J. *et al.* Puf Mediates Translation Repression of Transmission-Blocking Vaccine Candidates in Malaria Parasites. *PLoS Pathog.* (2013). doi:10.1371/journal.ppat.1003268
35. Miao, J. *et al.* Sex-Specific Biology of the Human Malaria Parasite Revealed from the Proteomes of Mature Male and Female Gametocytes. *Mol. Cell. Proteomics* (2017).
36. Stone, W. J. R. *et al.* Unravelling the immune signature of Plasmodium falciparum transmission reducing immunity. *Nat. Commun.* (2017).
37. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33,** 244–248 (2005).
38. Stone, W. *et al.* A Molecular Assay to Quantify Male and Female Plasmodium falciparum Gametocytes: Results From 2 Randomized Controlled Trials Using Primaquine for Gametocyte Clearance. *J. Infect. Dis.* **216,** 457–467 (2017).
39. Graves, P. M., Gelband, H. & Garner, P. Primaquine or other 8-aminoquinoline for reducing Plasmodium falciparum transmission. *Cochrane database Syst. Rev.* (2015). doi:10.1002/14651858.CD008152.pub4
40. Huynen, M. A., Snel, B. & Van Noort, V. Comparative genomics for reliable protein-function prediction from genomic data. *Trends in Genetics* (2004). doi:10.1016/j.tig.2004.06.003
41. Butler, C. L. *et al.* Identifying novel cell cycle proteins in apicomplexa parasites through co-expression decision analysis. *PLoS One* (2014). doi:10.1371/journal.pone.0097625
42. Treeck, M., Sanders, J. L. L., Elias, J. E. E. & Boothroyd, J. C. C. The Phosphoproteomes of Plasmodium falciparum and Toxoplasma gondii Reveal Unusual Adaptations Within and Beyond the Parasites' Boundaries. *Cell Host Microbe* **10,** 410–419 (2011).
43. Lasonder, E. *et al.* The plasmodium falciparum schizont phosphoproteome reveals extensive phosphatidylinositol and cAMP-protein kinase A signaling. *J. Proteome Res.* **11,** 5323–5337 (2012).
44. Pease, B. N. *et al.* Global analysis of protein expression and phosphorylation of three stages of plasmodium falciparum intraerythrocytic development. *J. Proteome Res.* **12,** 4028–4045 (2013).
45. Lasonder, E., Green, J. L., Grainger, M., Langsley, G. & Holder, A. a. Extensive differential

protein phosphorylation as intraerythrocytic *Plasmodium falciparum* schizonts develop into extracellular invasive merozoites. *Proteomics* n/a-n/a (2015). doi:10.1002/pmic.201400508

46. Bowyer, P. W., Simon, G. M., Cravatt, B. F. & Bogyo, M. Global profiling of proteolysis during rupture of Plasmodium falciparum from the host erythrocyte. *Mol Cell Proteomics* **10,** M110 001636 (2011).

47. Lindner, S. E. *et al.* Total and putative surface proteomics of malaria parasite salivary gland sporozoites. *Mol. Cell. Proteomics* **12,** 1127–43 (2013).

48. Okamoto, N., Spurck, T. P., Goodman, C. D. & McFadden, G. I. Apicoplast and mitochondrion in gametocytogenesis of Plasmodium falciparum. *Eukaryot. Cell* **8,** 128–132 (2009).

49. MacRae, J. I. *et al.* Mitochondrial metabolism of sexual and asexual blood stages of the malaria parasite Plasmodium falciparum. *BMC Biol.* **11,** 67 (2013).

50. Helb, D. A. *et al.* Novel serologic biomarkers provide accurate estimates of recent Plasmodium falciparum exposure for individuals and communities. *Proc. Natl. Acad. Sci. U. S. A.* 1501705112- (2015). doi:10.1073/pnas.1501705112

51. Cutts, J. C. *et al.* Immunological markers of Plasmodium vivax exposure and immunity: a systematic review and meta-analysis. *BMC Med.* (2014). doi:10.1186/s12916-014-0150-1

52. Crompton, P. D. *et al.* A prospective analysis of the Ab response to Plasmodium falciparum before and after a malaria season by protein microarray. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 6958–6963 (2010).

53. Schneider, P. *et al.* Submicroscopic Plasmodium falciparum gametocyte densities frequently result in mosquito infection. *Am. J. Trop. Med. Hyg.* **76,** 470–474 (2007).

54. Aurrecoechea, C. *et al.* PlasmoDB: A functional genomic database for malaria parasites. *Nucleic Acids Res.* (2009). doi:10.1093/nar/gkn814

55. Drakeley, C. J., Secka, I., Correa, S., Greenwood, B. M. & Targett, G. A. T. Host haematological factors influencing the transmission of Plasmodium falciparum gametocytes to Anopheles gambiae s.s. mosquitoes. *Trop. Med. Int. Heal.* (1999). doi:10.1046/j.1365-3156.1999.00361.x

56. Targett, G. *et al.* Artesunate reduces but does not prevent posttreatment transmission of Plasmodium falciparum to Anopheles gambiae. *J. Infect. Dis.* **183,** 1254–9 (2001).

57. Drakeley, C. J. *et al.* Parasite infectivity and immunity to Plasmodium falciparum gametocytes in Gambian children. *Parasite Immunol.* (2004). doi:10.1111/j.0141-9838.2004.00696.x

58. Sutherland, C. J. *et al.* Reduction of malaria transmission to Anopheles mosquitoes with a six-dose regimen of co-artemether. *PLoS Med.* **2,** 0338–0346 (2005).

59. Dunyo, S. *et al.* Gametocytaemia after drug treatment of asymptomatic Plasmodium falciparum. *PLoS.Clin.Trials* (2006).

60. Hallett, R. L. *et al.* Chloroquine/Sulphadoxine-Pyrimethamine for Gambian Children with Malaria: Transmission to Mosquitoes of Multidrug-Resistant Plasmodium falciparum. *PLoS Clin. Trials* (2006). doi:10.1371/journal.pctr.0010015

61. Ceesay, S. J. *et al.* Changes in malaria indices between 1999 and 2007 in The Gambia: a retrospective analysis. *Lancet* (2008). doi:10.1016/S0140-6736(08)61654-2

62. Shrestha, S., Li, X., Ning, G., Miao, J. & Cui, L. The RNA-binding protein PfPuf1 functions in the maintenance of gametocytes in *Plasmodium falciparum*. *J. Cell Sci.* jcs.186908 (2016). doi:10.1242/jcs.186908

63. Simon, N. *et al.* Sexual Stage Adhesion Proteins Form Multi-protein Complexes in the Malaria Parasite Plasmodium falciparum. *J. Biol. Chem.* (2009). doi:10.1074/jbc.M808472200

64. Alexa A and Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. *R package version 2.24.0.* (2016).

65. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6,** (2011).

66. van Dam, T. J., Wheway, G., Slaats, G. G., Huynen, M. A. & Giles, R. H. The SYSCILIA gold standard (SCGSv1) of known ciliary components and its applications within a systems biology consortium. *Cilia* **2,** 7 (2013).

642 67. Moreno-Pérez, D. A., Dégano, R., Ibarrola, N., Muro, A. & Patarroyo, M. A. Determining the
643 Plasmodium vivax VCG-1 strain blood stage proteome. *J. Proteomics* **113,** 268–280 (2015).
644
645

Acknowledgments

We are thankful to Colin Sutherland (LSHTM London, UK) for commenting on the manuscript and providing serum samples that we used to study antibody responses in Gambian individuals. David Conway and Johannes Dessens (LSHTM London, UK) gave helpful feedback on the initial gold standard lists. We further thank Adam D. Shandling, Jozelyn V. Pablo and Andy A. Teng from Antigen Discovery Inc. (Irvine, CA) for their work on the protein microarray.

TB was supported by the Netherlands Organization for Scientific Research (nwo.nl), through a VIDI fellowship 016.158.306. The Radboud Institute for Health Sciences (rihs.nl), supported LM through grant R-2765. The Virgo consortium (virgo.nl), grant FES0908, supported the work of RvdL and TJPvD. TWAK is supported by the Netherlands Organization for Scientific Research (NWO-VIDI 864.13.009). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions Statement

MAH, TB, RvdL and LMK conceptualised the work. LMK, RvdL, TJPvD, MAH and TB analysed the data. LMK, TB, DAB, PA, FS, CJJ, SMK, TWAK, MM and RS assembled the Gold Standard. KL, MvdVB, WG, RSS, LMK and WS conducted and analysed experiments with samples and resources provided by CD and JJC. LMK wrote the first draft of the manuscript, all authors reviewed the manuscript.

Competing financial interests

JJC is employed by Antigen Discovery, Inc. The authors declare no further competing interests.

678 **Figure Legends**

679

680

681 Figure 1***Figure 1. Clustered data sets used in this study with genes ranked according to***
682 ***their protein or transcript expression****. Level of expression as detected in the respective*
683 *samples with unique peptide counts for MS data and percentiles for transcriptomics. The*
684 *studies are clustered using complete linkage according to their overall gene expression*
685 *similarities (Euclidean distance). See Table 1 for study keys. Distribution of*
686 *asexual(a)/gametocyte(g) samples (red/blue) is shown in top bar, proteomics (P) and*
687 *transcriptomics (T) (dark/light grey) in lower bar.*

688

689 Figure 2***Figure 2. Gametocyte-specificity scores for P. falciparum genes derived from***
690 ***proteomics (P) and transcriptomics (T) data sets****. (A) Boxplot for integrated scores for*
691 *the two gold standard sets and all other Pf genes, derived from proteomics, transcriptomics*
692 *or all data sets (combined). (B) Density of P and T gametocyte scores, individual gold*
693 *standard genes and their scores are indicated at the bottom (red, asexual, blue gametocyte).*
694 *(C) 100 highest ranking proteins and transcripts, gametocyte gold standard in blue. (D)*
695 *Correlation of the gametocyte-specificity scores derived from all integrated MS studies and*
696 Pf *MS studies only.*

697

698 Figure 3***Figure 3. Validation of Bayesian gametocyte scoring with area under the curve***
699 ***(AUC) values****. Integrated data and individual data sets are compared by 10-fold cross-*
700 *validation (subsampling of gametocyte and asexual gold standard sets). Integrated*
701 *proteomics (P) and transcriptomics (T) scores in bold lines. P. berghei data sets in shades of*
702 *red, individual proteomics and transcriptomics studies with short and long dashes,*
703 *respectively. See Table 1 for study keys.*

704

705 Figure 4***Figure 4. Comparison of reported gametocyte-specific proteins in mass***
706 ***spectrometry studies.*** *(A) Proteins reported as gametocyte-specific by six individual*
707 *studies, agreements on gametocyte-specificity are summarized in the table. Bayesian:*
708 *gametocyte-specific proteins (n=602) that have a score > 5 after data integration. The*
709 *overlap with previously published data sets is shown, but not to scale. Overlap between the*
710 *individual studies is not shown for better visibility. Note that the Lasonder 2002 study*
711 *includes proteins that were found in gametocytes or gametocytes and gametes. (B) Proteins*
712 *that were reported as non-gametocytic and are (partially) included after data integration*

713

714 Figure 5***Figure 5. Seroprevalence in two cohorts of parasite carriers in The Gambia****.*
715 *(A+B) Antibodies against the highest scoring gametocyte-specific proteins were measured*
716 *on protein microarrays. Comparison of positivity (mixture-model cutoff) in gametocyte*
717 *carriers (n=164) and non-carriers (n=63). Gametocyte presence determined by microscopy.*
718 *All individuals were positive for asexual parasites. (A) Prevalence of antigens from the gold*
719 *standard (n=40) and predicted gametocyte-specific proteins (n=30), Mann-Whitney U test*
720 *(B) Antigens of five predicted gametocyte-specific proteins are preferentially recognized by*

721 *gametocyte carriers. Error bars indicate the upper limit of the 95% confidence interval*
722 *around the proportion. p<0.05 Fisher's exact test, corrected for multiple testing of a total of*
723 *70 antigens (Benjamini-Hochberg)*

724

725 Figure 6***Figure 6. Validation of gametocyte-specific targets in qRT-PCR*** *Targets are sorted*
726 *for decreasing gametocyte-specificity in all panels, see Table 2. (A) Minimum transcript*
727 *abundance in blood stage versus gametocytes in different Pf strains. 1000-fold enrichment*
728 *of transcript in gametocytes over asexuals was assumed when delta-Ct was 10 or higher*
729 *(dashed line), considering the lowest Ct value detected in any asexual concentration-*
730 *matched sample. This threshold was not met by the transcripts with gene IDs in grey. (B)*
731 *Detection limit of eight validated targets alongside Pfs25 in serial dilutions of Pf NF54*
732 *asexual stage parasites (ring stage parasites 10-20 hours post invasion). (C) Detection limit*
733 *of the most sensitive targets in serial dilutions of stage V gametocytes. (A-C) For Pf NF54,*
734 *all n=3, other strains n=2 biological replicates (error bars: standard error of the mean), all*
735 *measurements in triplicates. (D) Sensitivity of eight validated targets in Kenyan blood*
736 *samples of varying gametocyte densities.*

737

738 **Table 1. Data sets used for integration**

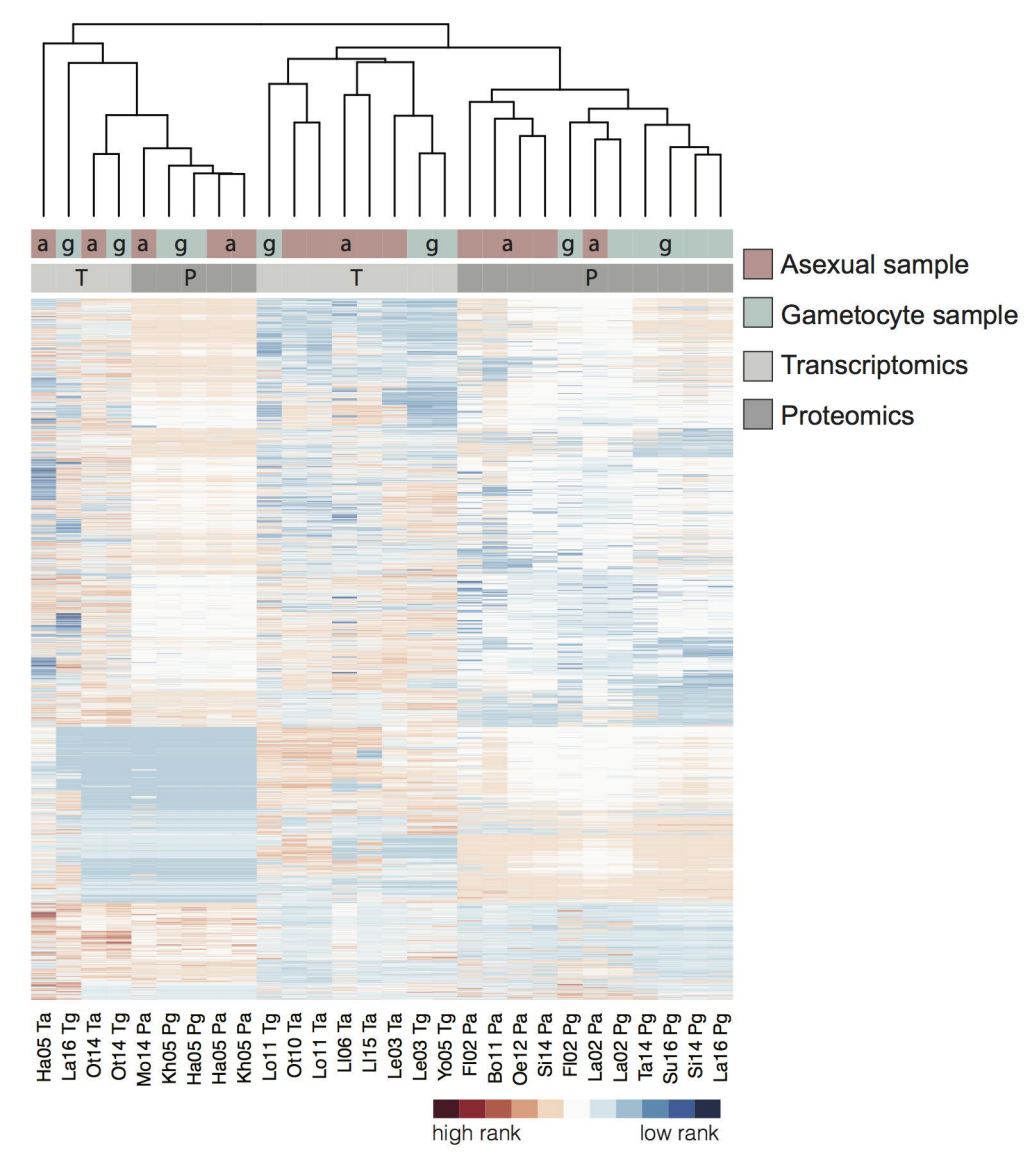
739 Life stages Spz sporozoites, Ri rings, Troph trohozoites, Schiz schizonts, Mer
740 merozoites, Gct gametocytes. MS mass spectrometry (highest unique peptide count
741 in any of the samples), T transcriptomics (highest percentile in any of the samples).
742 Data from Miao et al. 2017 was integrated after analyses of high scoring proteins,
743 ranks and scores are included in Supplementary Table 2.
744 *Asexual microarray data by Llinas and others retrieved from plasmoDB version 28
745 (Data set "Pfal3D7 real-time transcription and decay"), no accompanying publication.

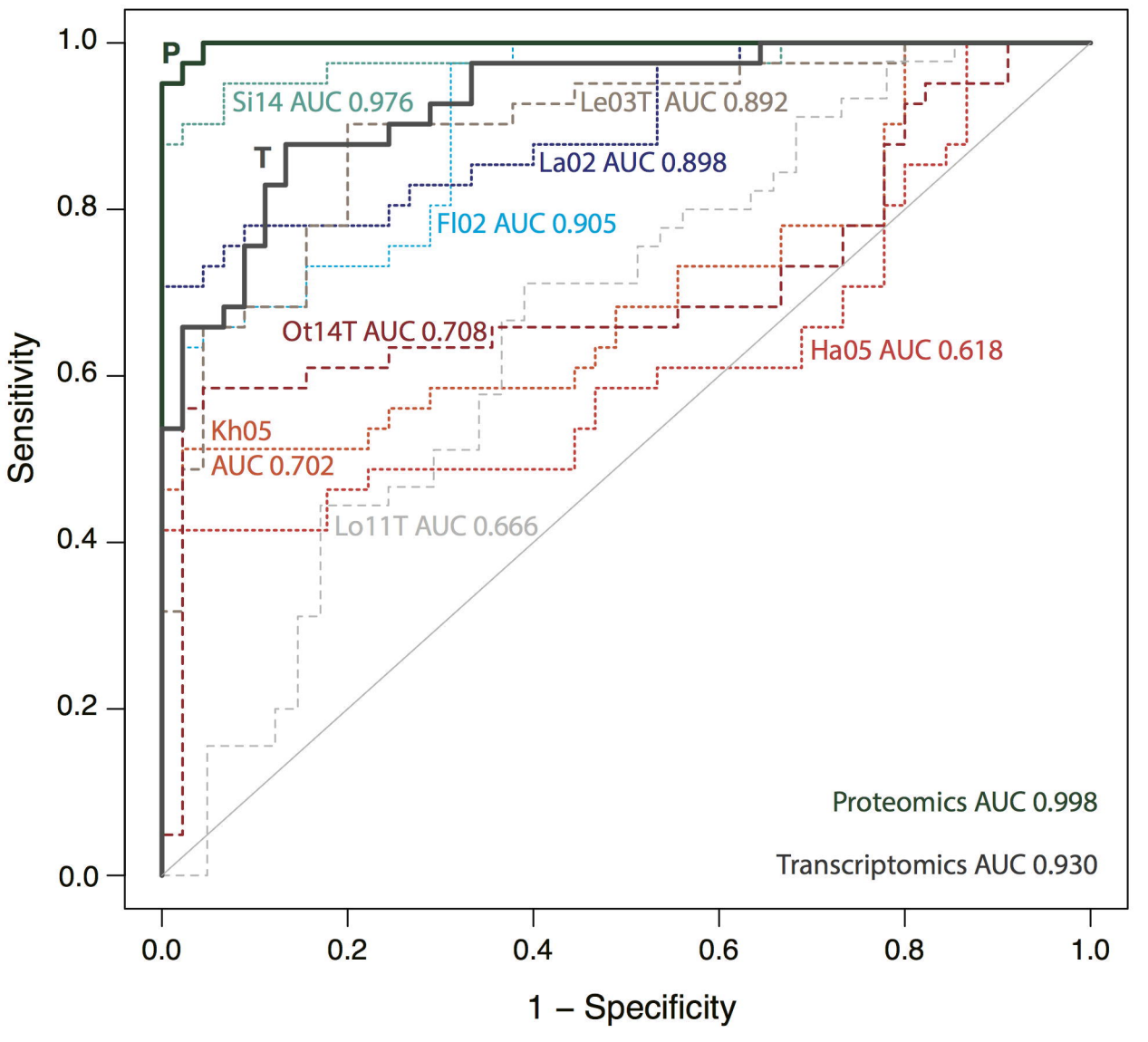
Study	Reference	Species	Life Stage	Integrated Data
Fl02	Florens et al. 2002 [14]	*Pf*	Spz, Troph, Mer, Gct	MS asexual/Gct
La02	Lasonder et al. 2002 [13]	*Pf*	Troph, Schiz, Gct	MS asexual/Gct
Le03	LeRoch et al. 2003 [23]	*Pf*	Spz, Ri, Troph, Schiz, Mer, Gct	T asexual/Gct
Ha05	Hall et al. 2005 [15]	*Pb*	Ri, Troph, Schiz, Gct	MS + T asexual/Gct
Kh05	Khan et al. 2005 [16]	*Pb*	Mixed blood stage, Gct	MS asexual/Gct
Yo05	Young et al. 2005 [24]	*Pf*	Gct	T Gct
Ll06	Llinas et al. 2006 [25]	*Pf*	All blood stages, synchronized	T asexual
Ot10	Otto et al. 2010 [26]	*Pf*	All blood stages, synchronized	T asexual
Bo11	Treeck et al. 2011 [42]	*Pf*	Schiz	MS asexual
Lo11	Lopez-Barragan et al. [27]	*Pf*	Troph, Schiz, Gct	T asexual/Gct
Oe12	Oehring et al. 2012 [18]	*Pf*	Ri, Troph, Schiz	MS asexual
Mo14	Moreno-Perez et al. 2014 [67]	*Pv*	Ri, Troph, Schiz	MS asexual
Ot14	Otto et al. 2014 [28]	*Pb*	Ri, Troph, Schiz, Gct	T asexual/Gct
Si14	Silvestrini et al. 2010, Tao et al. 2014 (re-analyzed) [17,19]	*Pf*	Troph, Schiz, Gct	MS asexual/Gct
Ta14	Tao et al. 2014 [19]	*Pf*	Gct	MS Gct
Ll15	Llinas et al. 2015*	*Pf*	All blood stages, synchronized	T asexual
La16	Lasonder et al. 2016 [20]	*Pf*	Gct	MS + T Gct
Su16	Suarez-Cortes et al. 2016 [21]	*Pf*	Gct	MS Gct
Mi17	Miao et al. 2017 [35]	*Pf*	Gct	MS Gct

746
747
748

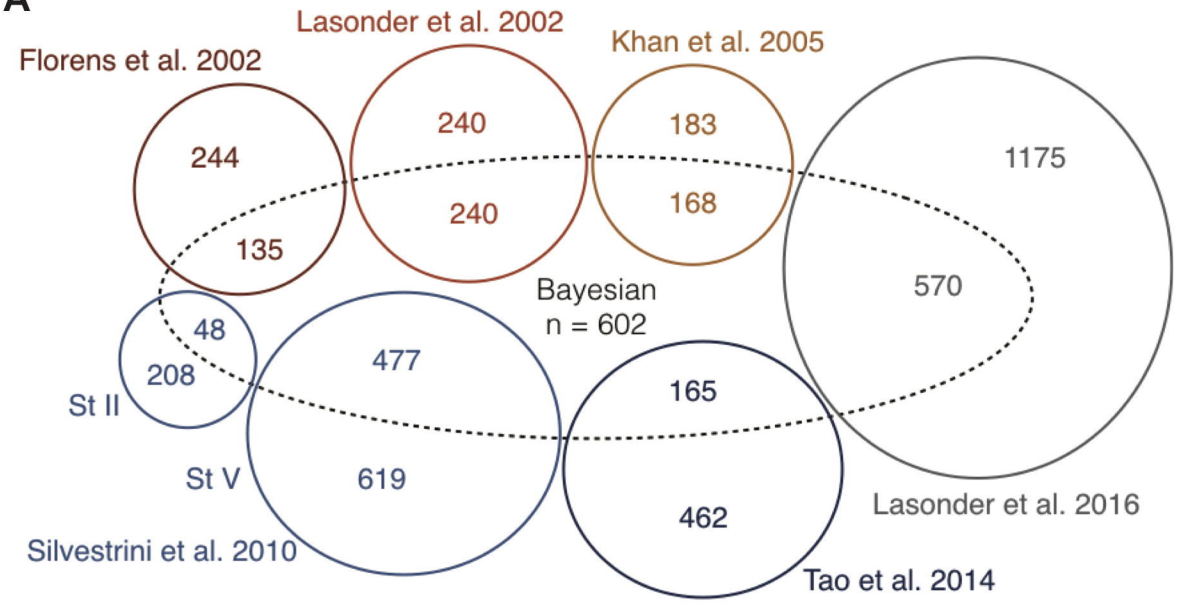
Rank	Gene ID	Description	Name	Intron-spanning	Least Ct difference
3	PF3D7_1143600	conserved Plasmodium protein, unknown function	–	no	9.37
5	PF3D7_1147200	tubulin--tyrosine ligase, putative	–	no	8.52
6	PF3D7_1026100	conserved Plasmodium protein, unknown function	–	yes	12.94
7	PF3D7_1438800	conserved Plasmodium protein, unknown function	–	yes	8.57
8	PF3D7_0625100	sphingomyelin synthase 2, putative	SMS2	no	11.04
12	PF3D7_0930000	procollagen lysine 5-dioxygenase, putative	–	no	11.96
14	PF3D7_0518700	mRNA-binding protein PUF1	PUF1	yes	8.31
15	PF3D7_0303900	phosphatidylethanolamine-binding protein, putative	–	yes	10.87
16	PF3D7_1466600	conserved Plasmodium protein, unknown function	–	no	5.67
17	PF3D7_1107900	mechanosensitive ion channel protein, putative	MSCS	no	6.33
18	PF3D7_1214500	conserved Plasmodium protein, unknown function	–	yes	10.61
24	PF3D7_1131500	conserved Plasmodium protein, unknown function	–	no	10.41
51	PF3D7_0929600	G2 protein, putative	–	yes	8.37
61	PF3D7_0816800	meiotic recombination protein DMC1, putative	DMC1	yes	13.29
75	PF3D7_0903800	LCCL domain-containing protein	CCp4	yes	12.70

Table 2: Properties of putative gametocyte-specific targets.
Rank in transcriptomics (all data sets) for specificity in gametocytes. Random sample of top 100, excluding the gold standard. If primers are not intron-spanning, samples were DNase I treated. Ct difference is the difference between the lowest Ct detected in asexual samples and the highest Ct in concentration-matched stage V gametocytes, averaged across strains *Pf* NF54, NF135, NF166 and NF175.


a g a g a g a g a g a g
T P T P
Asexual sample
Gametocyte sample
Transcriptomics
Proteomics
Ha05 Ta
La16 Tg
Ot14 Ta
Ot14 Tg
Mo14 Pa
Kh05 Pg
Ha05 Pg
Ha05 Pa
Kh05 Pa
Lo11 Tg
Ot10 Ta
Lo11 Ta
Ll06 Ta
Ll15 Ta
Le03 Ta
Le03 Tg
Yo05 Tg
Fl02 Pa
Bo11 Pa
Oe12 Pa
Si14 Pa
Fl02 Pg
La02 Pa
La02 Pg
Ta14 Pg
Su16 Pg
Si14 Pg
La16 Pg
high rank
low rank

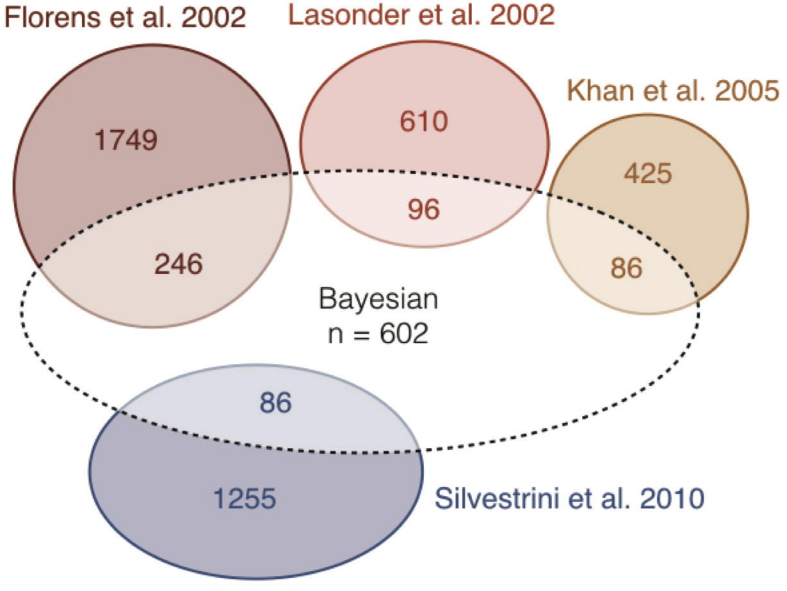

P
T
Si14 AUC 0.976
Le03T AUC 0.892
La02 AUC 0.898
Fl02 AUC 0.905
Ot14T AUC 0.708
Ha05 AUC 0.618
Kh05
AUC 0.702
Lo11T AUC 0.666
Proteomics AUC 0.998
Transcriptomics AUC 0.930
Sensitivity
1 – Specificity
0.0
0.2
0.4
0.6
0.8
1.0

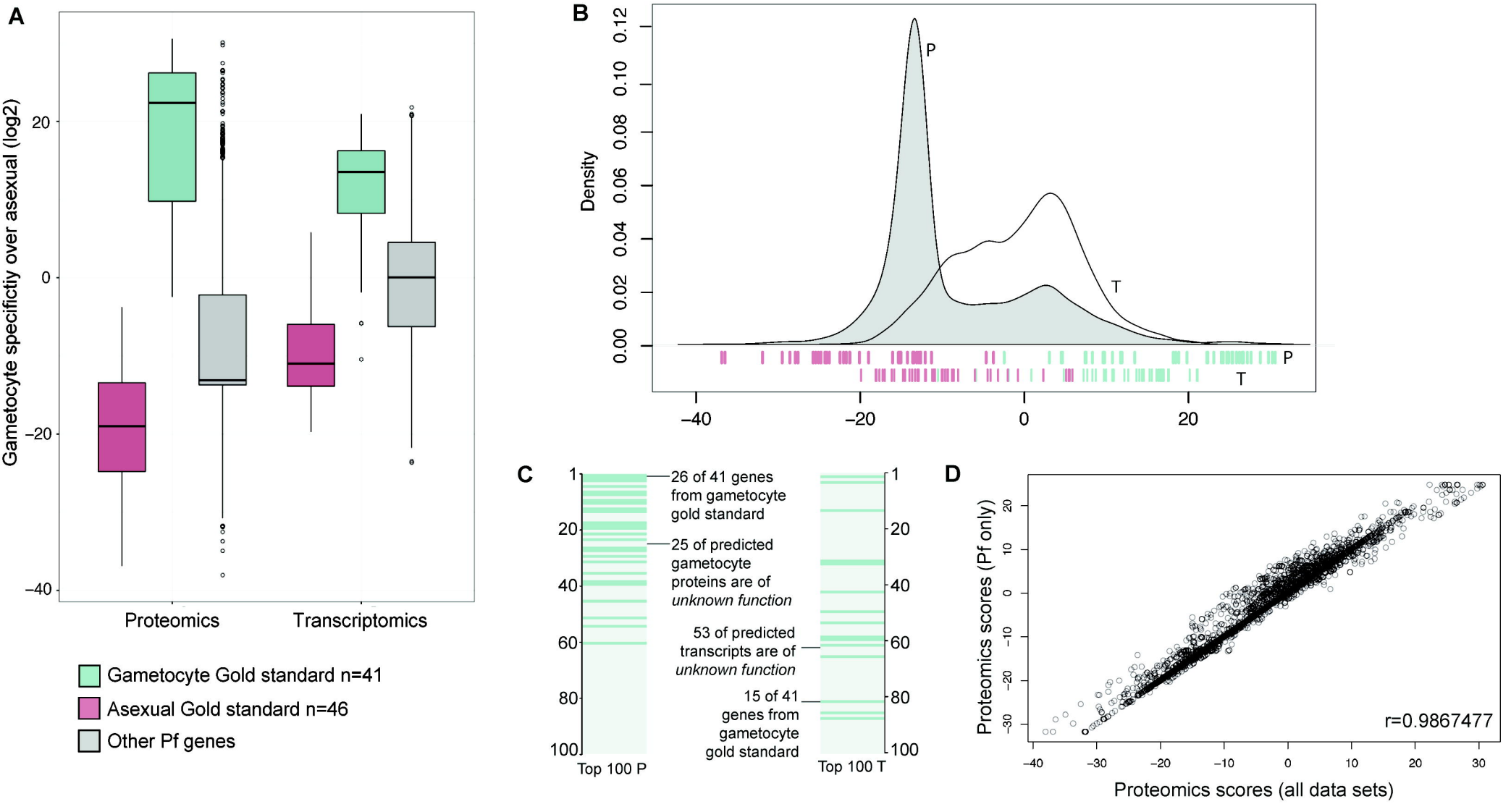
A

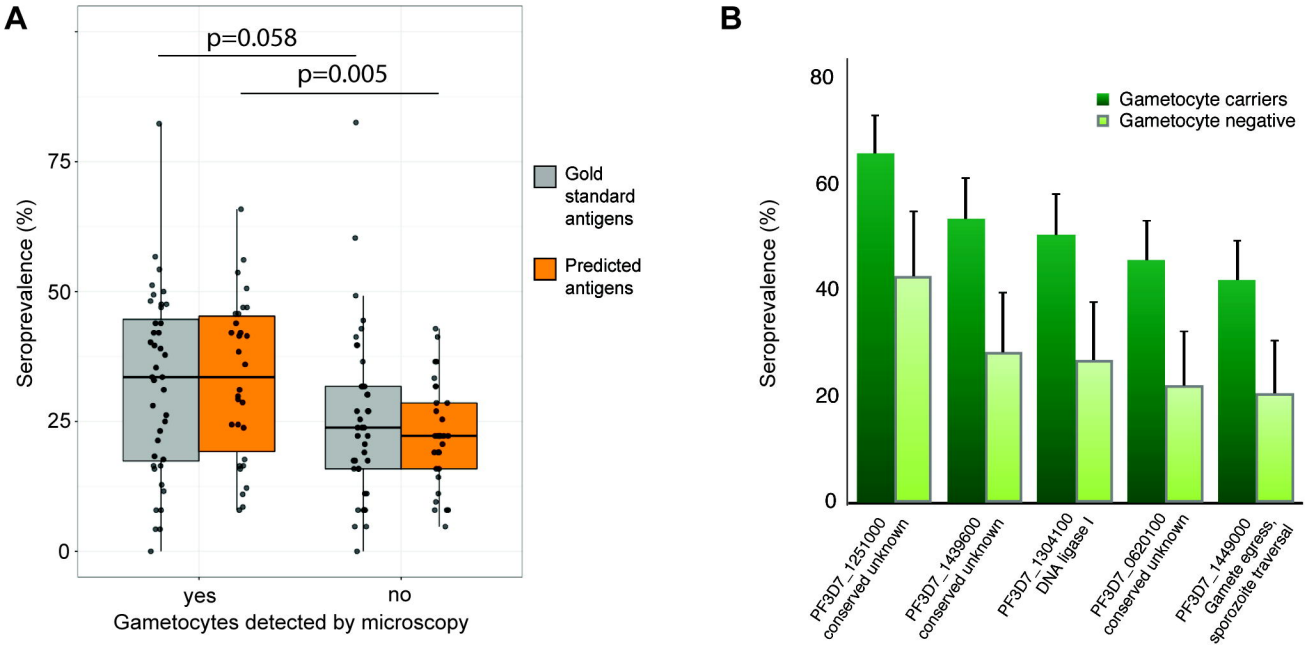


studies reporting as gametocyte-specific or enriched	number of proteins
0	7
1	48
2	156
3	209
4	117
5	51
6	14
total	602

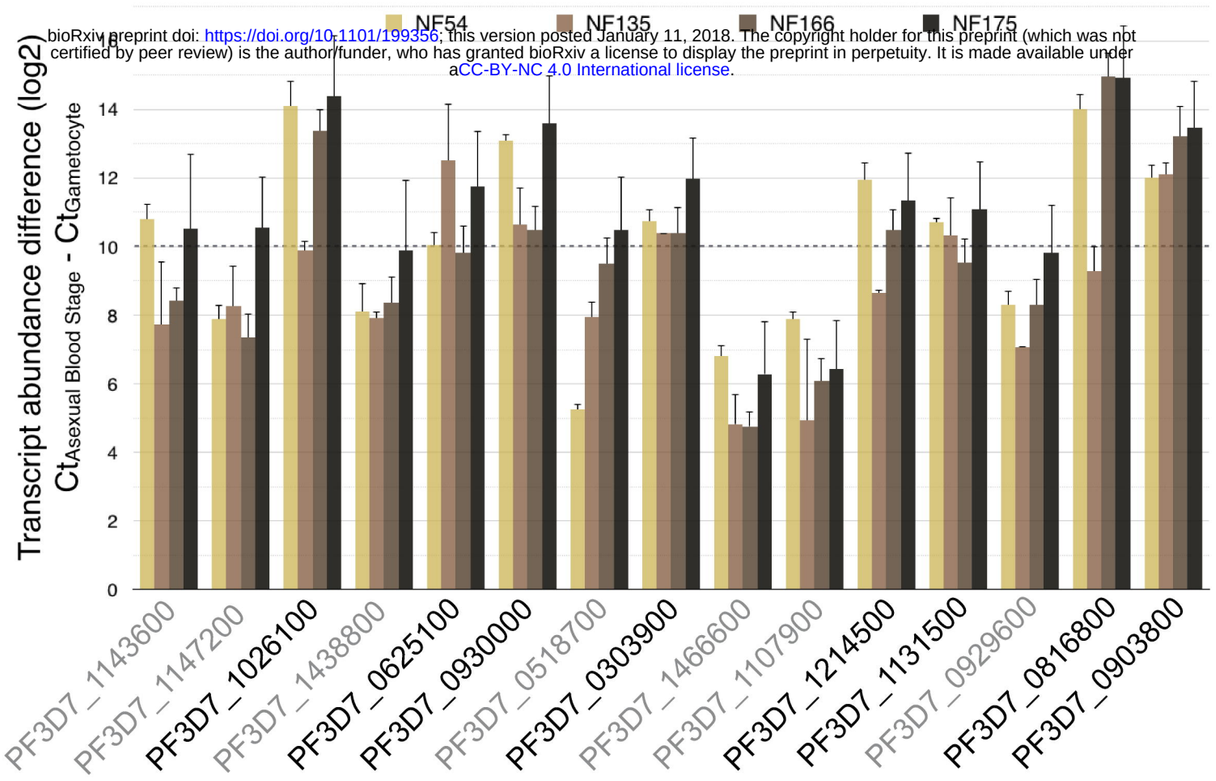
B



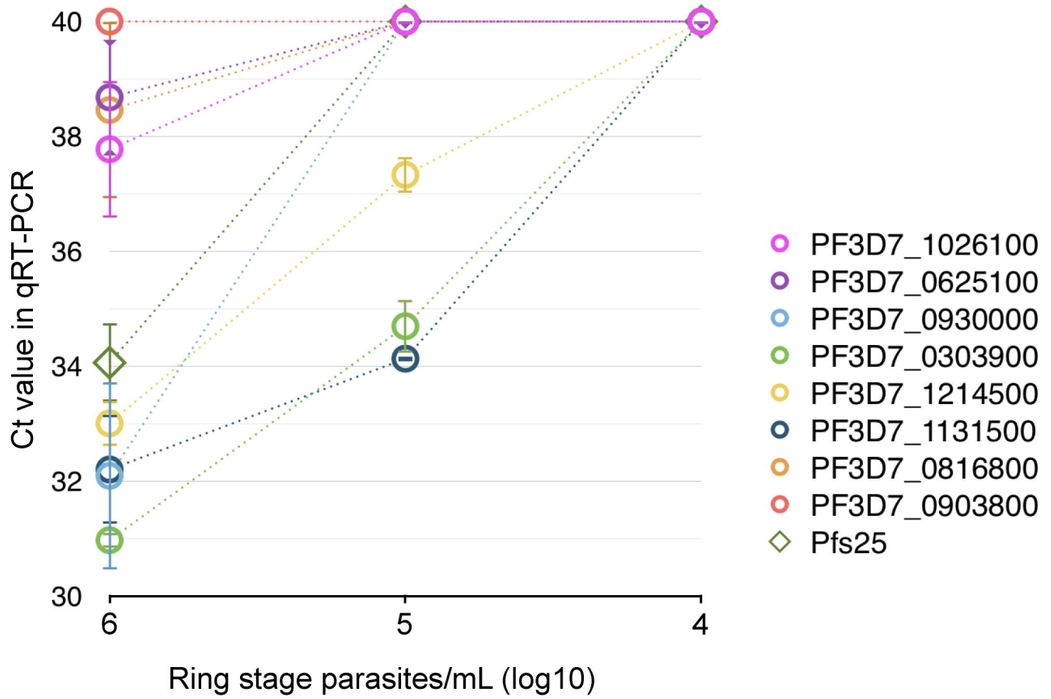

A
Gametocyte specifictiy over asexual (log2)
20
0
−20
−40
Proteomics
Transcriptomics
Gametocyte Gold standard n=41
Asexual Gold standard n=46
Other Pf genes
B
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
P
T
−40
−20
0
20
C
1
20
40
60
80
100
26 of 41 genes from gametocyte gold standard
25 of predicted gametocyte proteins are of unknown function
53 of predicted transcripts are of unknown function
15 of 41 genes from gametocyte gold standard
Top 100 P
Top 100 T
D
Proteomics scores (Pf only)
Proteomics scores (all data sets)
r=0.9867477
−40
−30
−20
−10
0
10
20
30


A
p=0.058
p=0.005
Seroprevalence (%)
75
50
25
0
yes
no
Gametocytes detected by microscopy
Gold standard antigens
Predicted antigens
B
Seroprevalence (%)
80
60
40
20
0
Gametocyte carriers
Gametocyte negative
PF3D7_1251000 conserved unknown
PF3D7_1439600 conserved unknown
PF3D7_1304100 DNA ligase I
PF3D7_0620100 conserved unknown
PF3D7_1449000 Gamete egress, sporozoite traversal

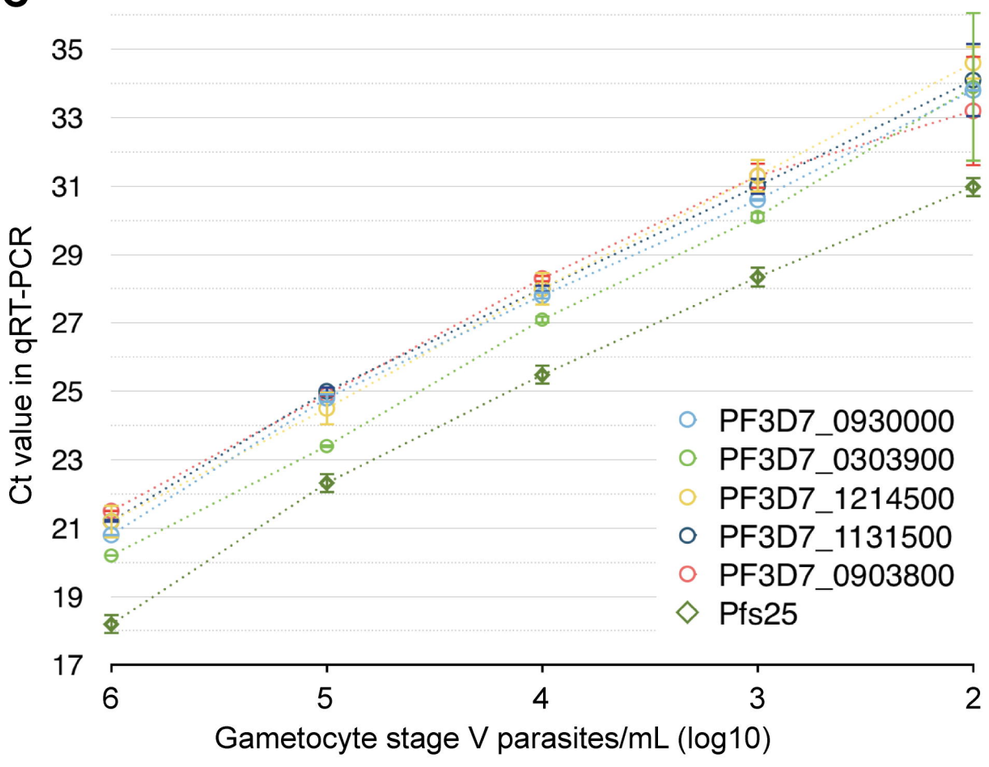
A



B



C



D