*Nature*. Author manuscript; available in PMC 2014 September 13.

Published in final edited form as:
*Nature*. 2014 March 13; 507(7491): 253–257. doi:10.1038/nature12970.

# A cascade of DNA binding proteins for sexual commitment and development in *Plasmodium*

**Abhinav Sinha**[#1], **Katie R. Hughes**[#1], **Katarzyna K. Modrzynska**[#2], **Thomas D. Otto**[2], **Claudia Pfander**[2], **Nicholas J. Dickens**[1], **Agnieszka A. Religa**[1], **Ellen Bushell**[2], **Anne L. Graham**[1], **Rachael Cameron**[1], **Bjorn F.C. Kafsack**[3], **April E. Williams**[3,4], **Manuel Llinas**[3,4,5], **Matthew Berriman**[2], **Oliver Billker**[2,†], and **Andrew P. Waters**[1,†]

[1]Wellcome Trust Centre for Molecular Parasitology, University of Glasgow

[2]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

[3]Department of Molecular Biology, Princeton University, Princeton, New Jersey, USA

[4]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, USA

[#] These authors contributed equally to this work.

## Abstract

Commitment to and completion of sexual development are essential for malaria parasites (protists of the genus *Plasmodium*) to be transmitted through mosquitoes[1]. The molecular mechanism(s) responsible for commitment have been hitherto unknown. Here we show that PBAP2-G, a conserved member of the ApiAP2 family of transcription factors, is essential for the commitment of asexually replicating forms to sexual development in *P. berghei*, a malaria parasite of rodents. PBAP2-G was identified from mutations in its encoding gene, PBANKA_143750, which account for the loss of sexual development frequently observed in parasites transmitted artificially by blood passage. Systematic gene deletion of conserved ApiAP2 genes in *Plasmodium* confirmed the role of PBAP2-G and revealed a second ApiAP2 member (PBANKA_103430, termed PBAP2-G2) that significantly modulates but does not abolish gametocytogenesis indicating that a cascade of ApiAP2 proteins are involved in commitment to the production and maturation of gametocytes. The data suggest a mechanism of commitment to gametocytogenesis in *Plasmodium* consistent



[†]Corresponding authors: OB (OB4@sanger.ac.uk) APW (Andy.Waters@glasgow.ac.uk) .
[5]Currently: Department of Biochemistry and Molecular Biology and Center for Infectious Disease Dynamics, The Pennsylvania State University, State College PA, USA

**Author Contributions:** APW and OB directed the research. AS generated the GNP clones performed some of the EMSA analyses, made pbap2-g gene knock out lines and complementation lines and analysed the latter. KRH performed microarray analyses, generated reporter and minigene constructs, made transgenic parasites and analysed them, performed competition experiments. KKM made the complementation construct, generated and analysed knock out and complemented lines for pbap2-g and pbap2-g2 and performed and analysed competition and microarray experiments. CP generated ko lines for pbap2-g and pbap2-g2 and performed the initial parasitological analysis. EB generated recombinase engineered constructs for use at WTSI and UoG. ALG and AAR performed expression analyses. NJD performed statistical analyses of motif distribution and assisted with the microarray analyses. RC performed the complementation experiments and transmission experiments. AEW performed EMSA analyses and generated constructs used in the analysis. TDO and MB generated the GNP sequence data and SNP analyses. ML and BK performed microarray analyses, ML and AW performed EMSA analyses and generated recombinant PB PBAP2-G DBD. APW, OB, AS, KRH and KKM wrote the paper.


Europe PMC Funders Author Manuscripts

with a positive feedback loop involving PBAP2G which might be exploited to prevent the transmission of this pernicious parasite.

Malaria parasites spontaneously and stochastically produce sexual forms (gametocytes) required for mosquito transmission. Asexual parasites commit to sexual development in the erythrocyte and the cell cycle-arrested male and female gametocytes are available to initiate transmission when ingested within the blood meal of a female anopheline mosquito. Gametocyte production may be lost when *Plasmodium* parasites are maintained either in continuous culture or by blood transfer between vertebrate hosts[1]. We generated three gametocyte non-producer (GNP) lines (GNPm7, GNPm8 and GNPm9) that had verifiably lost the ability to undertake gametocytogenesis after 52 weeks of mechanical passage (Fig. 1A, Fig. S1, Table S1).

Subsequent developmental stages (gametes, ookinetes) were absent and none of the GNP lines could be transmitted through mosquitoes (Fig. S2, Table S2). Whole genome sequencing (WGS) of these and an existing GNP line (ANKA 2.33) revealed numerous SNPs and indels per line (Fig. S3, Table S3), however, only a single gene, PBANKA_143750, carried a different and therefore independent nonsense or missense mutation in each line (Fig. 1B). PBANKA_143750 (or *pbap2-g*) encodes a putative transcription factor (TF) predicted to be composed of 2330 amino acids (aa) with a single 55 aa AP2 class DNA binding domain (DBD) at its C-terminus (Fig. 1B). PBAP2-G belongs to the 27 strong[4,5] *Plasmodium* ApiAP2 family of TFs themselves part of the larger Apetala 2 / Ethylene Response Factor (AP2/ERF) family of TFs restricted to the Plantae and apicomplexan protists. The role of PBAP2-G in gametocyte production was confirmed either by correcting the mutations in *pbap2-g* in the GNP lines through genomic recombination with a wt copy (generating GNPm7REP, GNPm8REP, GNPm9REP & 2.33REP respectively) or genetic complementation of a targeted deletion mutant of *pbap2-g* (Fig. 1C, Fig. S4A-G). Functionality of the restored gametocytes was demonstrated in GNPm7REP and 2.33REP by transmission through mosquitoes (Fig. 1D, Table S4). Disruption of a second ApiAP2 gene, PBANKA_103430 (*pbap2-g2*) (Fig. 1B) resulted in the nearly complete (>95%) loss of mature gametocytes, but in contrast to *pbap2-g*$^{-}$ parasites, small numbers of female gametocytes were occasionally observed (Fig. 1C). These were not, however, transmitted successfully to mosquitoes. In direct growth competition assays *pbap2-g*$^{-}$ parasites outgrew wt *P. berghei* and *pbap2-g2*$^{-}$ parasites, which had wt growth rates (Fig. 1E, S5). *pbap2-g*$^{-}$ mutants are therefore uniquely capable of converting a loss of gametocytes into increased asexual growth, which confers an advantage during asexual growth and explains why continued blood passage invariably selects for mutations in *pbap2-g*. This demonstrates that PBAP2-G functions specifically at the point of commitment, while PBAP2-G2 is required downstream, once sexual differentiation has become irreversible (Fig. 1E).

In a protein binding microarray the recombinant DBD of PBAP2-G[6,7] recognised closely related DNA motifs (Fig. 2A, Table S5) identical to the previously derived motif for the DBD from the orthologous ApiAP2 protein of *P. falciparum* (PF3D7_1222600)[6] confirming that both DBDs bind primarily to the same (GxGTACxC) motif. EMSA analyses (Fig. 2A)




Europe PMC Funders Author Manuscripts

refined the motif to two 6mers (GxGTAC and GTACxC, which are essentially palindromes of each other) that are sufficient and necessary for binding. A single point mutation in the core GTAC was sufficient to abrogate binding (Fig. 2A). These two motifs occurred within 2kb upstream of 49% of all genes (2359 of 4803 considered), yet more frequently in genes designated as upregulated in gametocytes (246 (54%) of 452 genes hypergeometric $P < 0.002$). The occurrence of both motifs upstream of *pbap2-g* itself suggested the potential for an autoregulatory feedback mechanism, and the regions of the genome containing these motifs upstream of *pbap2-g* were both recognised by PBAP2-G in EMSA analysis (Fig. 2A). Expression analysis demonstrated transcription of *pbap2-g* in blood stage parasites, however, epitope tagging of full-length *ap2-g* produced no detectable protein (Fig S6) yet gametocytogenesis was unaltered implying tagged PBAP2G activity is unaffected. However, a truncated CFP-tagged transgene product could be detected in nuclei of female gametocytes (Fig. 2C, S7).

Comparative microarray analyses showed that gametocyte specific genes were highly enriched among the 500 most down regulated genes in GNP lines ($p < 10^{-51}$; Fisher's exact test), *pbap2*-deletion parasites ($p < 10^{-74}$) and in the *pbap2-g2* deletion mutant ($p < 10^{-49}$) although less marked in the latter (Tables 1, S6). Comparison of the transcriptomes of wt asexual blood stage parasites with those of various *pbap2-g*$^{-}$ lines was performed in an attempt to identify early-transcribed genes downstream of and under control of PBAP2-G (Fig. 3A). The steady state transcription levels of 307 genes were identified as being down regulated (>2 SD reduced from the mean, Table S6) in schizonts.

The activity of 18 promoters consistently down regulated in GNP lines, and which contain one or more candidate PBAP2-G binding motifs, was analysed in wt and GNPm9 parasite backgrounds. Male, female or sex specific genes downstream of AP2G in the gametocyte developmental pathway were identified (Fig. 3B, S8, Table S8). Single point mutations in PBAP2-G binding motifs did not significantly reduce stage- or sex-specific expression of all of a number of reporter genes *in vivo*, even if identical changes ablated DNA binding *in vitro*. Only larger promoter truncations produced an impact on expression (Fig. S9). Therefore, the relatively simple and highly abundant PBAP2-G motif is only active in context and its presence not always indicative of a critical role for the activity of a particular promoter. The PBAP2-G motifs upstream of *ap2-g* itself do appear to be important as gametocytogenesis is blocked when the allelic motifs are both deleted supporting the concept that commitment to gametocytogenesis requires a positive feedback loop powered by PBAP2-G itself (Fig. 3C).

The discovery of the ApiAP2 family[4] was the first identification of predicted TFs in apicomplexan genomes, otherwise thought to be remarkably lacking in genes encoding TFs[8]. The majority of ApiAP2 TFs are probably essential, involved in the progression of the intraerythrocytic asexual development of *Plasmodium*. Roles for additional ApiAP2 factors in the continuation of development of parasite forms associated with transmission was demonstrated, namely the ookinete (PBAP2-O[9]), sporozoite (PBAP2-S[10]) and liver stage (PBAP-L[11]) development. ApiAP2 proteins may also silence genes, possibly through maintenance of heterochromatin[12]. The AP2/ERF family in *Plasmodium* are predicted to act singly or in combinations that control the continuation of the transcriptional programme of

the *Plasmodium* life cycle[4,6]. Heritable gene regulatory strategies include epigenetic marks, stable cytoplasmic factors, and transcriptional autoregulatory circuits that can determine distinct cell fates[13]. In the latter, commitment to a specific developmental pathway (e.g. gametocytogenesis) is probabilistic, its frequency being defined by the likelihood of the interaction of a fate-determining TF with a critical promoter often triggering a positive autoregulatory feedback loop that commits the cell[14], a paradigm that has been invoked within the *Plasmodium* AP2 TF network[6]. *P. falciparum* uses precise epigenetic control to influence the sub-nuclear location of *pfap2-g*[15] and therefore possibly PFAP2-G binding which when coupled to an autoregulatory positive feedback loop (Fig. 3C) involving PBAP2-G production, could provide flexible control of gametocytogenesis in a manner that would also be amenable to environmental sensing[16, 17]. AP2-G is to date unique within the apiAP2 TF family in that it directs a change in developmental fate rather than merely progressing a lineage (Fig. S10) distinguishing it from AP2-G2 and from a number of other genes required for gametocyte maturation[18]. This critical role of AP2-G is conserved in *P. falciparum* (Kafsack et al.) even though models for the timing of commitment in the two parasites differ[19,20]. Orthologues of the *ap2-g* DBD are present in all sequenced *Apicomplexa* raising the possibility that mechanisms of commitment to sexual development may also be conserved (Fig. S9). Thus these data identify the earliest known event in parasite transmission. Since it occurs in the blood of the host it is amenable to and suggests novel control strategies largely through drug development and offers some strategic value in the prevention of sexual development and reduction of transmission.

## Materials & Methods

### Parasite lines and Methods

*P. berghei* ANKA HP was obtained from Chris Janse at Leiden University Medical Centre and was originally referred to as clone 15Cy1A (Leiden Malaria Group web site). Line 820 was generated from HP. *P. berghei* ANKA clone 2.33 is a non-gametocyte producing cloned line reported in 1990 and is now widely distributed and grown by mechanical passage[23]. All infections were performed on female TO mice (age 6-8 weeks; weight 25-30g) according to home office licence regulations and the local ethical committees. All animals were assigned to experiments without pre-selection and no blind assignations were performed. Serial passage of freshly cloned *P. berghei* reference line 820 m1cl1[3] was performed as follows: 10 mice (m1-m10) were initially infected with 200 μl of a 1:200 dilution of a mouse infected with line 820 at a parasitaemia of ~2%. In the absence of any a priori information concerning mutation rates in P. berghei a sample size of 10 was selected based on concerns of animal welfare, cost and logistics. Each week the infections were passaged to a further 10 mice in a similar manner when the parasitaemia was >1%. Parasitaemia and gametocytaemia were monitored by examination of Giemsa-stained blood films and by flow cytometry as described[24-28]. The infected blood from each mouse was also cryopreserved each week. Passage to a fresh mouse was halted when a line was negative for gametocyte production for 4 consecutive weeks and designated GNPmx where x would be 1-10. The experiment was halted after 52 weeks. Lines GNPm7, 8 & 9 were cloned by limiting dilution, clones subjected to negative selection[29] to remove the selectable marker residual in the GFP:RFP selection cassette and cloned once more. Each parasite cloning procedure employed 10 mice

and mice were infected by i.v. tail injection with an average of 1.5 parasites which in our experience will give rise to 4 infected mice. Negative selection involved 3 mice the infections of which were assayed by PCR for completeness of selection. Lines generated in this way were designated m(7,8,9)mxClx indicating the mouse and clone number identifiers from the negative selection process. In the main text these cloned negatively selected lines are simply referred to GNPm7, 8 and 9.

Transfection of GFP- and RFP-expressing 'wild type' parasites from the *P. berghei* line 820 with linearised targeting constructs, selection and cloning of the mutant parasites were performed according to procedures previously described[30]. Genotypic analysis of transfected parasites was performed by Southern analysis of FIGE separated chromosomes and by diagnostic PCR on genomic DNA. Details of the primers used for PCR are shown in Table S9. Phenotype analysis of mutant parasites during blood stage development, quantification of gametocyte production and ookinete development *in vitro* was performed using standard methods as previously described[25-28]. Mosquito stage development was analysed in *Anopheles stephensi* mosquitoes using standard methods of infection of mosquitoes and analysis of oocyst and sporozoite production and analysis of sporozoite infectivity to TO mice[31]. The capacity of wild type and engineered parasites to infect mice by mosquito interrupted feeding was determined by exposure of female TO mice (n = 2–4) to 40–50 mosquitoes, at day 21 after the infectious blood meal. Infection was monitored by analysis of blood stage infection in Giemsa stained films of tail blood at day 4 till day 8 after infection. Infectivity was recorded as 'wild type' if mice developed a parasitemia of 0.1–0.5% at day 4 after infection. For the 2.33 rescue experiment images representative of >80 gametocytes at a parasitaemia of 8.2 % and gametocytaemia of 5.4 % are shown in Figure 1D; similar results were seen on three consecutive days.

## DNA-Sequencing

To sequence clones 2.33, 820, GNPm7, GNPm8, and GNPm9, libraries of 300-500 bp fragment length were generated following a PCR-free protocol[32]. The libraries were sequenced using an Illumina Genome Analyser II with the V4 chemistry. Summary of reads for each project including accession codes are given in Table S3. Data are available at http://www.ebi.ac.uk/ena/data/view/ERP000253

## Sequencing - *De novo* assembly

We generated a *de novo* assembly of reads from the 820 parental clone using with velvet[33] version 1.0.12 and the following parameters: -exp_cov auto -min_contig_lgth 500 -cov_cutoff 10 -ins_length 350 -min_pair_count 20. We obtained 417 supercontigs with an N50 of 240kb. We processed the assembly as described in the post-assembly genome-improvement toolkit protocol[34]. In short, scaffolds were ordered with ABACAS[35] against the *P. berghei* ANKA reference genomes (GeneDB, version July 2010). This resulted in 16 pseudomolecules (14 chromosomes and 2 plastids) and a 'bin' of 100 contigs that could not be associated with a chromosome. Next, using scaffolds of at least 1 kb as a substrate, IMAGE[36] was used to close 469 (61 %) of the 774 sequencing gaps. Single base and small insertion or deletion (indel) errors were corrected using ICORN[37]. This corrected 1067 single-base errors and 92 indels. 1589 positions had heterozygous calls, which represented

collapsed repeats, mostly in BIR genes. Last, the annotation of the *P. berghei* ANKA reference genome was transferred onto the improved *P. berghei* 820 assembly using RATT[38] (Assembly option). In total 4821 of the 4938 gene models were transferred correctly.

The assembly is available on ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/berghei/820/.

## Sequencing - Variant calls

To call variants, SMALT (version 0.6.2 http://www.sanger.ac.uk/resources/software/smalt/, parameters: −r 0; −x; −y 0.8; −i 1000; and for index a k-mer size of 17 (−k) and a step size of 3 (−s)) was used to map reads against the generated 820 assembly. After generating bam files with the SAMtools package[39], variation was called with GATK[40] (parameters: -ploidy 1-glm POOLBOTH-pnrm POOL). For the reads mapped onto the 820 assembly, the variation of each clone, and concordance with other clones was analysed using a PERL script. For the reads mapped onto the ANKA reference genome, the script ignored variants that were called in all m7-m9 clones as well as 820. The quality filter for a variant was 60. The whole WGS and SNP analytical pipeline is summarised in Fig. S12.

Variant calling in *Plasmodium* from resequencing data is inherently noisy, due to false calls within repeats and low complexity regions. Thus 3 independent clones were used to identify coincident site(s). Isolate-specific variation is catalogued in Table S3 and the large proportion of heterozygous calls are highlighted (a manifestation of calling variants within repetitive and low complexity regions). All data were generated using ad hoc scripts (available upon request). The variant (.vcf) files of the each isolate are available from ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/berghei/820/vcf.

## Phylogenetic analysis

Data were generated from the results of a BLASTP search of EuPathDB apicomplexa using the AP2 domain from PBANKA_143750 as the query. Significant hits were defined as those that covered at least 75% of the length of the query domain and had >50% conserved residues. Neighbor Joining tree generated in CLC Genomics Workbench (v6.5.1) using the Jukes-Cantor protein distance measure. Values shown are for 1000 bootstrap iterations. The tree is rooted using the most distant *Arabidopsis thaliana* DBD Q9M0l0.2.

## Recombinant Protein production

N-terminal glutathione S-transferase (GST) fused extended ApiAP2 DNA binding domains (cloned into pGEX-4T1) from *P. falciparum ap2-g* (PFL_1085w) and *P. berghei ap2-g* (PBANKA_143750) were expressed in Rosetta (DE3) pLys S competent cells with 0.2 mM IPTG at 25 °C and batch-purified using affinity chromatography (Glutathione HiCap Matrix slurry; Qiagen). The purity of protein was estimated by 10% SDS PAGE and the eluted proteins were quantified with spectrophotometry by optical absorbance at 260 nm. The eluted protein yield was concentrated and buffer exchanged using (Amicon Ultra-0.5 Centrifugal filter devices; 30K device; Millipore). The properties of the DBD fusion proteins produced and used in this study are indicated in Table S10.

### Protein binding microarray (PBM) analysis

PBMs were processed and analysed as previously described[41-43] Quoting from Campbell et al[41] "Briefly, custom designed oligonucleotide arrays are double-stranded using a universal primer, incubated with GST-AP2 fusion proteins, visualized with Alexa-488 conjugated anti-GST antibody, and scanned using an Axon 4200A scanner. Proteins were used at the maximum concentration obtained from purification and represent one-fifth of the total reaction volume used on the PBM. In this study three different universal platforms were used covering all contiguous 8-mers as well as gapped 8-mers spanning up to 10 positions. After data normalization and calculation of enrichment scores the "Seed-and-Wobble" algorithm was applied to combine the data from two separate experiments and create position weight matrices (PWMs). An enrichment score cut-off of 0.45 was used to distinguish high affinity binding data from low affinity and nonspecific binding. The score for each 8-mer reflects the affinity of a DNA binding domain for that sequence, with higher scores representing tighter interactions. Secondary motifs were identified by running the "rerank" program until E-scores below 0.45 were obtained." Furthermore, the motifs are a graphical representation of each position weight matrix and are visualized using enoLOGOS[44].

### EMSA - DNA binding of purified N-terminal GST fusions of AP2 domains of AP2-G of *P. falciparum*

(PF3D7_1222600) and *P. berghei* (PBANKA_143750) to their cognate DNA sequences was analysed by electrophoretic mobility shift assay (EMSA). Single stranded oligonucleotides containing the recognition motif flanked either by random nucleotides (same for all flanking sequences) or by the actual genome sequence (as they occur naturally in the 5′ upstream regions of potential AP2 target genes) and their corresponding complementary oligonucleotides were synthesised and purchased from MWG Eurofins (Germany) as labelled (5′-biotinylated & HPLC purified) and unlabelled sequences. Complementary single-stranded oligonucleotides were annealed to create double-stranded probes and used for EMSA as labelled and unlabelled target probes for DNA-binding domain of AP2G. EMSAs were performed using the LightShift Chemiluminescent EMSA kit (Pierce). Briefly, 2μg of the purified GST fusion of PFAP2-G and PBAP2-G (in separate reactions) was pre-incubated with 0.02pmol of the labelled probe in 20μl the binding reaction containing binding buffer, 1 μg poly(dI-dC), 50% glycerol, 100 mM $MgCl_2$, 1% NP40, and 60 μg BSA at room temperature for 10 minutes. The unlabelled probe (4 pmol; 200-fold excess to the labelled probe) was then added as a competitor and the reaction was incubated for further 20 minutes at room temperature. The reaction was fractionated using 12% PAGE and transferred to a nylon membrane (Hybond) as per manufacturer's instructions. Specific binding of the AP2 domain with the target motif was detected as an upward shift using the Chemiluminescence Nucleic Acid Detection Module (Pierce) as per the manufacturer's instructions and anti-GST antibodies.

### Southern blot analysis

Southern blot analysis from wild type line 820 and three different *ap2-g* length-variable knockouts was performed to show successful integration of the selectable marker cassette at

the desired genetic locus. Briefly, approximately 10 μg of Plasmodipur (EuroProxima) filtered and purified genomic DNA from lines 820 (WT), G401cl1 (complete orf knockout), G418cl6cl3 (DNA-binding domain knockout) and G529cl2 (partial orf knockout bearing the GNPm7, 8 and 9 mutations) was double-digested each with 7 μl of appropriate restriction enzyme (New England Biolabs) pairs at 37°C for 4 hours with NEB Buffer4. For comparison with the WT line (820), gDNA from WT and G401cl1, WT and G418cl6cl3 & WT and G529cl2 was double-digested with the High-Fidelity (HF) versions of NcoI & SpeI, NcoI & BamHI and EcoRI & SpeI, respectively. After transfer the membrane was hybridized (60°C overnight) with $P^{32}$ labelled single-stranded DNA probe for a specific region from one of the homology arms used for generating the gene targeting vector. The probes were PCR-amplified and purified using the following oligonucleotides: GU1058/ GU1059 for G401cl1, GU1416/GU1417 for G418cl6cl3 and GU1414/GU1415 for G529cl2. The membrane was washed thrice with decreasing concentration of SSC (3× SSC, 1× SSC, 0.5×SSC) and exposed to a maximum resolution X-ray film (BioMax MR film; Kodak) for 35 hours.

## Northern blot analysis

Approximately 5 μg of RNA sample for each line (except G529cl2; which was ~2 μg) was denatured and fractionated in 1.2% agarose gel in 2.2 M (w/v) formaldehyde at 20 V overnight in 1× MOPS as running buffer. After transfer the RNA in the membrane was hybridized (60°C overnight) with $P^{32}$ labelled single-stranded DNA probe for *p28* mRNA (PBANKA_051490; 0.62kb ORF) and normalised using *hsp70* mRNA probe (PBANKA_071190; 2.08kb ORF), washed and exposed to a maximum resolution X-ray film (BioMax MR film; Kodak).

## Recombineering Methods

Gene knock out vectors for *pbap2-g* and *pbap2-g2* were submitted to the *Plasmo*GEM database as PbGEM-072446 and PbGEM-039238, respectively[21] where details of their construction can be found. Complementation vectors were made using the Red recombination system of phage *lambda* using published protocols [45] First *E. coli* harbouring *P. berghei* gDNA clone PbG01-2472c01, which carries a >11 kb genomic insert including *pbap2-g*, were rendered competent for recombination by transfection with plasmid pSC101gbaA[46]. A marker cassette for positive and negative selection in *E. coli*, attR1-zeo-pheS-attR2, was then amplified using primer pairs Comp143750UpR1/2 or Comp143750D1R1/2 (see supplemental Table S9 for primer sequences). The resulting PCR products carried 50 bp extensions homologous to the upstream or downstream intergenic regions of *pbap2-g*, respectively. The PCR products were introduced into the recombination competent *E. coli* carrying the PbG01-2472c01 library plasmid and the recombination product selected with zeocin. The bacterial marker was then exchanged for the *P. berghei* selection marker *hdhfr-yfcu* in an *in vitro* Gateway reaction, the product of which was retransformed into *E. coli* and negatively selected on YEG-Cl and kanamycin as described[45]. Clones carrying the correct complementation plasmid were identified by PCR across the boundary of the *hdhfr-yfcu* cassette. Before transfection the constructs were linearized using NotI removing the plasmid backbone.

### Reporters – construct generation

The CFP reporter construct pG0148 was generated by inserting CFP into pG073 as follows: CFP was amplified from pL1382 using primers to incorporate XhoI and SmaI restriction sites. This was cloned into the XhoI / SmaI sites of pG073 (KH unpublished) between an *hsp70* (PBANKA_071190) promoter (1.4 kb) and *p45/48* constitutive 3′ UTR. The plasmid also contains a negative selection cassette[29] and target regions for DXO integration into a *p230p* locus downstream of the GFP/RFP cassette in the 820 line[28]. Candidates for reporter analysis in the first batch (rep 1 – 14) were chosen based on fold down regulation in GNP vs 820 schizont, the presence of at least one predicted AP2 binding motif (GTAC.C, OR G.GTAC OR G.GTAC.C) and at least moderate expression levels in at least one life cycle stage. For some of the second batch of reporters based on analysis of trophozoite stage transcripts (rep 15 – 24) the additional criteria of not predicted to be translationally repressed was included. 2kb of sequence immediately upstream to the predicted translational start site (PlasmoDB) was amplified by PCR using Taq polymerase and primers incorporating KpnI/XhoI restriction sites. pG0148 was digested with KpnI/XhoI to excise the *hsp70* promoter and new reporter promoters ligated in. To introduce mutations into the predicted AP2-G binding sites an overlapping PCR strategy was used to mutate the GTAC to GTAA. A primer designed around the site incorporating the mutation in both forward and reverse complement was used with the original forward and reverse primers for the 2kbp fragment in a 2 stage overlapping PCR reaction. The fragment was cloned into pG0148 and sequenced to confirm the mutation. After verification of correct insert 15 – 30 μg of plasmid DNA was digested with SacII to linearise the integration fragment and subsequently cut with either ScaI or SapI to cut the plasmid backbone and minimise risk of introducing episomes. Fully digested DNA was ethanol precipitated and resuspended in water before being mixed with 100 μl Nucleofector (Lonza Amaxa) solution for transfection into 820 and GNPm9 lines.

### Reporters – Transfection

DNA prepared as above (4 μg – 12 μg per transfection) was mixed with Nycodenz purified synchronous *Plasmodium berghei* schizonts lines 820 or GNPm9 and electroporated using programme U33 of Amaxa machine. Parasites were then immediately injected into the tail vein of a TO mouse. 24 – 28 hours following transfection the parasites were placed on positive selection by including pyrimethamine (Sigma) in drinking water[30].

### Reporters – Flow Cytometric Analysis

Analysis was performed on parasites from tail blood on days 6 – 10 after transfection. 2 μl of tail blood was placed into 500 μl rich PBS (PBS (Roche) with 20 mM Hepes, 20 mM Glucose, 4 mM $NaHCO_3$, 0.1% BSA) containing 1 μl Vybrant DyeCycle Ruby (Invitrogen) and incubated at 37°C for 30 min. Parasites were pelleted and resuspended in 1.5 ml of FACS buffer (PBS (Roche) with 2 mM Hepes, 2 mM Glucose, 0.4 mM $NaHCO_3$, 0.01% BSA, 2.5 mM EDTA). Analysis was performed on a CyAn ADP 9 colour flow cytometer (Beckman Coulter) equipped with 405 nm, 488 nm and 642 nm solid state lasers and 500,000 events were acquired (counting all events except debris). On each day an uninfected control and CFP negative parental controls were processed in parallel with reporter lines.

Europe PMC Funders Author Manuscripts

Data analysis was performed using Kaluza analysis software (Beckman Coulter) following the gating strategy indicated in the following schematic. For histogram analysis the CFP geometric mean expression level (AFU) in each gated population male, female and asexual was calculated as a mean from three day's data and plotted as a bar chart in excel.

All events were plotted Forward scatter (FS) vs Side scatter (SS) and gate E drawn to exclude debris. Events in gate E were plotted on FSvsFS (area) and gate J(1) drawn to exclude potentially autofluorescent doublets and clumps. Events in gate J(1) were plotted FS vs Ruby (DNA stain) and gate G drawn to select infected cells. Gate G was drawn based on a negative (uninfected) control population stained in the same way and analysed on the same occasion (Fig. S13a).

Events in gate G were plotted SS vs CFP and a CFP positive gate drawn based on a non-CFP expressing parental line (820, HP or GNP9) stained and processed on the same occasion and at similar parasitaemia. GFP vs RFP was plotted for all infected cells (events in G) and for only those falling into the CFP+ve gate. Gates drawn on female F (RFP positive) and male M (GFP+ve) populations was used to calculate the percentage of each population which express CFP based on the number of cells in each gate in each plot. For illustrative figures the infected population (G) was plotted on GFP vs RFP and those additionally falling into gate CFP+ve coloured magenta while those not CFP+ve coloured grey. The percentage of the population within each gate that was expressing CFP (calculated as above) is indicated (Fig. S13b).

## Microscopy Analysis

For some lines the CFP expression was analysed on a Zeiss Axioplan II fluorescent microscope. A drop of tail blood was stained with 5 μM Hoechst in enriched PBS for 10 min then placed on a microscope slide under a coverslip and sealed with nail varnish and visualised under a 100× oil immersion objective, images were captured and processed using Volocity software.

## Methods for promoter interruption experiments

During attempts to rescue gametocytogenesis in GNP lines by complementation rescue techniques we had observed that an interruption to the AP2-G promoter slightly downstream of two G.GTAC motifs led to a loss of gametocyte production. To investigate this further a series of constructs was made to target the AP2-G endogenous promoter and mutate specifically in the region of these G.GTAC motifs. Effect on gametocytogenesis after integration of these constructs into the endogenous AP2-G promoter in the fluorescent 820 parental line could then be monitored using flow cytometry.

## Promoter interruption construct generation and transfection

A double crossover homologous recombination method was used to create targeted interruptions of the *ap2-g* endogenous promoter. The plasmid pL0035 was used which contains a selection cassette including Human dhfr driven by *pbeef1aa* promoter surrounded by multiple cloning sites. Genomic fragments from the *ap2-g* promoter region were amplified by PCR from wild type genomic DNA using Kapa Hi-Fi polymerase

(KapaBiosystems) and cloned in piecewise as described below to allow for flexibility with the vector for creating multiple mutations. The 207bp region containing the G.GTAC motifs was synthesised by MWG with or without point mutations in the core motif. All regions are described by their distance from the *ap2-g* gene start. A downstream integration fragment from bp-416 to bp-1277 was cloned in using SmaI and EcoRI and an upstream integration region from bp-2695 to bp-1912 cloned in using HindIII and SacII. The region from bp-1913 to bp-1484 was cloned downstream of the selection cassette and in front of the downstream integration region using KpnI and EcoRV to create vector pG266 (2kb del). Using SmaI and EcoRV the synthesised region from −1913 to −1484 either wild type or containing single point mutations in the G.GTAC motif was cloned into vector pG266 to create pG298 (2kb WT) or pG312 (2kb MutA). Additionally a clone containing the WT 200 bp region in reverse orientation was selected pG299 (2kb WT Rev). Subsequently the SmaI cloning site in pG298 was removed to created pG313 (2kb WT-Sma). To extend the region of endogenous promoter remaining between the selection cassette and the *ap2-g* gene an additional fragment from −2870 to −1913 was cloned into the KpnI site downstream of the selection cassette in pG266 and pG313 to create pG266+3 (3kb del) and pG313+3 (3kb WT). Constructs were linearised using HindIII and EcoRI, and approximately 10 μg of purified linear DNA were transfected in to *P. berghei* parasites (820 line) as described elsewhere.

## Promoter interruption Gametocytogenesis Assays

Gametocyte levels in transfected parasites were monitored by flow cytometry (on a FACS CyAN, Beckman Coulter) on a drop of tail blood from animals containing the transfected parasites and maintained on pyrimethamine selection throughout from 6 days post-transfection for up to 5 consecutive days. Parasites were passaged into a clean animal maintained on pyrimethamine selection and gametocytaemia followed. Since the background gametocyte levels measurable using our methods in the parental 820 line varied from ~3 to ~20% depending on parasitaemia and unknown factors, a control transfection was carried out to enable gametocyte levels to be monitored in a line that had been maintained under exactly the same conditions. This was usually the plasmid pG306 which integrated to the *p230p* locus and contains a CFP gene driven by the PBANKA_101870 promoter. This also enabled us to confirm general transfection efficiency in each batch of transfections. After gating on the infected population using DyeCycle Ruby staining the percentage of parasites expressing RFP (female) or GFP (male) parasites was calculated. Results shown are the total gametocytaemia (male and female) as a percentage of the parasite population and a mean +/− SD from three readings from passaged animals. The 820 parental line is a mean from four readings.

## Minigene construction and analysis

pG0148 was generated as previously described in reporters section. To generate pG0157 a 2kb fragment immediately upstream of the *ap2-g* gene was amplified using primers to incorporate KpnI and XhoI restriction sites and cloned in place of the *hsp70* promoter in pG0148. To generate pG0189 a 300bp fragment of *ap2-g* was amplified to incorporate XhoI restriction sites and was cloned in frame with CFP into the XhoI restriction site between the HSP70 promoter and the CFP gene in pG0148. To generate pG0190 CFP was amplified

from pL1382 using primers to exclude the Stop codon of CFP and incorporate XhoI and SmaI restriction sites. This was cloned into pG073 to generate pG0188 (not shown). A 900 bp C-terminal fragment of *ap2-g* incorporating the DNA binding domain was amplified from genomic DNA using primers to incorporate SmaI restriction sites and cloned into the SmaI restriction site downstream of and in frame with CFP in pG0188. To generate pG0191 the *ap2-g* promoter and first 300bp of coding sequence were amplified using primers incorporating KpnI and XhoI restriction sites and was cloned in place of the *hsp70* promoter in pG0190. Plasmids were sequenced and 5-10 μg of linearised purified DNA transfected into either 820 or GNPm9 lines as previously described for reporter genes. Resulting transfected parasites were analysed by flow cytometry and fluorescence microscopy for expression and localisation of CFP signal. Each experiment was performed independently three times

## Competitive growth assays

GNPm9M1Cl1 was transfected with construct pG0148 to constitutively express CFP from an *hsp70* promoter to generate line GNP-CFP. An analogous construct with RFP driven by the *hsp70* promoter was generated (pG0161) and transfected into Wild type (HP) producer line to generate WT-RFP. Also generated was a wild type (HP) producer line expressing CFP from construct pG0148 (WT-CFP). Each line was individually grown in a TO mouse under pyrimethamine selection. 2 μl tail blood from each mouse was stained with Vybrant Dyecycle Ruby (Invitrogen) to label infected RBCs and then run on a CyAn ADP 9 Colour flow cytometer (Beckman coulter). After gating on infected cells the CFP or RFP expression was analysed showing that nearly 100% of each population after gating for infected cells expressed the fluorescent marker. Parasites were mixed to create a 50:50 mix of parasites containing either WT-CFP and WT-RFP or GNP-CFP and WT-RFP. These were injected (IV) into mice. Parasites were monitored daily by flow cytometry and after gating for infected cells the % of the population expressing either RFP (gate AF−+), CFP (gate AF+−) or both (gate AF++) reflecting mixed-multiply infected cells was calculated and plotted. On day 6, blood from each mouse was passaged into a new host and the time course continued.

After day 11 parasites were cryopreserved. For the competition assays between the *ap2-g* KO1, *ap2-g2* KO and *p28* KO, the *Plasmo*GEM KO vectors were transfected into the GFP- and mCherry expressing parasites. Once the parasitaemia in transfected animals reached ~5%, they were used to generate an inoculum containing an equal proportion of red and green parasites. Accuracy of each inoculum was tested using flow cytometry. New mice were injected ($1\times10^5$ parasites per animal) and kept under continued pyrimethamine treatment to prevent the emergence of untransfected parasites. The proportion of red and green parasites in the mixture was followed daily using flow cytometry. Three infected mice were used for each comparison.

## Microarray Methods

A 8X15k custom microarray (Agilent) providing coverage of the *P. berghei* genome at >1 probe per kbp of coding sequence was used[22]. Samples were prepared from parasites maintained using standard parasitological procedures. For schizont cultures parasites were obtained from cardiac puncture and grown overnight in culture. For ring stage cultures

parasites were matured in vitro to schizont stage in order to synchronise the population, then injected into a new host and allowed to reinvade. Blood was harvested at 24 +6 hpi and filtered through a magnetic column (variomacsD) to deplete of mature stages and gametocytes. For Trophozoite stage parasites parasites prepared as for ring stages were then cultured for a further 6 hours. All samples were filtered through a Plasmodipur filter to remove mouse leucocyte contamination before RNA preparation using a standard trizol method. Samples were processed for microarray using methods as described[22]. For GNP and *ap2-g* KO1 a two-colour microarray hybridisation was performed with a background pool of cDNA made from material from all life cycle stages (except late mosquito and liver stages). Parental control lines and experimental samples were then hybridised with the same background pool sample for all experiments. For *ap2-g* KO2 and *ap2-g* KO, the mutant samples were hybridised against the equivalent samples from the parental line and against each other. Arrays were scanned on an Agilent Microarray Scanner. Normalized intensities were then extracted using Agilent Feature Extractor. All expression data are available from the Gene Expression Omnibus database (www.ncbi.nih.gov/geo) under the accession numbers GSE52859 and GSE53246.

### Statistical methods for microarrays

Three biological replicates were performed for each life cycle stage of *ap2-g* KO1 line and the 820 parental line. Naturally derived GNP line (schizonts only) microarray results are representative of two technical replicates each from three independently derived GNP lines. These technical replicates were performed in different laboratories using the same methods. The *ap2-g* KO1 and GNP microarray data was uploaded to PUMADB (http://puma.princeton.edu/) for further processing. The data was extracted as a log(2) of the fold change of red (sample) vs. green (common pool) with minimal filtering to exclude background signal and median centred. The fold change between the GNP sample and the 820 parental line was calculated for each transcript and the mean and standard deviation of the replicates calculated (using Microsoft Excel). The distribution of these samples was confirmed to be normal ($P<2.2\times10\text{-}16$ Kolmogorov-Smirnov test in R version 2.10), and the transcripts classed as down regulated in GNP lines were those 2 SD below the mean fold change. For plotting volcano plots (fig 3) a two tailed t-test was performed on the independent replicates and a −log10 transform of this result plotted. This was plotted against the log(2) fold change using R ggplot2 library To determine which transcripts were gametocyte-specific the fold change between three replicates of gametocyte stage wild type parasites was compared to three replicates of schizont stage wild type parasites. A one tailed t-test was then used to determine those upregulated in gametocytes as highlighted in volcano plots in fig 3. For *ap2-g* KO2 and *ap2-g*2 the biological triplicates of each of the hybridisations (both mutants against the WT and against each other) were processed using the R v 2.15.0 software[47] with limma package[48]. The data was background corrected and normalized between the arrays (LOESS normalisation). Fold changes between the strains and p-values for differential expression were calculated with a linear statistical model. The p-values from all experiments were adjusted using the false discovery rate (FDR) correction.

For the gametocyte expression rank (Fig. 3A and Table S6) the absolute intensity values from microarrays from three independent replicates of wild type gametocytes was used and

ranked from highest (1) to lowest (~4553) expression rank. In order to test for the deregulation of the gametocyte specific genes in all the strains, the enrichment in gametocyte specific genes (expression rank 1 to 500) in the top 500 genes showing the highest fold change in each of the mutants was tested using the Fisher's exact test. Comparisons of the variances of the microarray data were carried out in R and all the variances were similar; none of the samples were significantly different (P<10E-16, F-test). Microarray data has been submitted to the GEO database (Accession Numbers: GSE52859 and GSE53246).

### Search for DNA binding motifs

The genomic sequences for all *P. berghei* genes were identified using PlasmoDB (version 9.1) and defined as a 2kb region upstream of the transcription start site to the first base of the transcription start site (4803 entries). A file was also created for the gametocyte-specific genes (452 entries). Differences in usable entries were due to genes close to the ends of chromosomes or poorly assembled regions, and regions that overlapped other genes. A custom Perl script was used to count occurrences of the PBAP2-G and PBAP2-G2 motifs in the sequences using a regular expression (PBAP2-G was defined as /G.GTAC|GTAC.C/ and PBAP2-G2 was defined by orthology as /TGC.ACC|GGT.GCA/ Campbell et al, (2010)). The script counts the occurrence of each pattern per-region and also provides a total number of sequences that contain at least one occurrence, and is available on request. Hypergeometric p-values were calculated interactively using R version 2.10.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

The authors declare they have no competing interests. We thank Angela McBride for technical assistance with mosquito transmissions; Ana Rita Gomes, Jaishree Tripathi, Diane Vaughan and the III flow cytometry facility for assistance; Clare Cairney and Nicol Keith at the Beatson Institute for use of their Agilent microarray scanner; Alfred Cortes and David Baker for critical reading of the manuscript.

APW is funded by the Wellcome Trust (Ref. 083811/Z/07/Z). AS was a student of the University of Glasgow Wellcome Trust 4-year PhD Programme Molecular Functions in Disease. APW, OB and MB are members of Evimalar (Ref. 242095) which funds the work of TDO. Work at the Sanger Institute was funded by grants from the Wellcome Trust (098051) and the Medical Research Council (G0501670). ML is funded by NIH R01 AI076276 and the Centre for Quantitative Biology (P50GM071508). BFCK was supported by a HHMI fellowship of the Damon Runyon Cancer Research Foundation.

## Reference List

1. Janse CJ, et al. Plasmodium berghei: in vivo generation and selection of karyotype mutants and non-gametocyte producer mutants. Exp. Parasitol. 1992; 74(1):1. [PubMed: 1730264]
2. Mair GR, et al. Universal features of post-transcriptional gene regulation are critical for Plasmodium zygote development. PLoS. Pathog. 2010; 6(2):e1000767. [PubMed: 20169188]
3. Ponzi M, et al. Egress of Plasmodium berghei gametes from their host erythrocyte is mediated by the MDV-1/PEG3 protein. Cell Microbiol. 2009; 11(8):1272. [PubMed: 19438517]
4. Balaji S, et al. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. Nucleic Acids Res. 2005; 33(13):3994. [PubMed: 16040597]

5. Painter HJ, Campbell TL, Llinas M. The Apicomplexan AP2 family: integral factors regulating Plasmodium development. Mol. Biochem. Parasitol. 2011; 176(1):1. [PubMed: 21126543]
6. Campbell TL, et al. Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. PLoS. Pathog. 2010; 6(10):e1001165. [PubMed: 21060817]
7. Berger MF, Bulyk ML. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat. Protoc. 2009; 4(3): 393. [PubMed: 19265799]
8. Gardner MJ, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature. 2002; 419(6906):498. [PubMed: 12368864]
9. Yuda M, et al. Identification of a transcription factor in the mosquito-invasive stage of malaria parasites. Mol. Microbiol. 2009; 71(6):1402. [PubMed: 19220746]
10. Yuda M, et al. Transcription factor AP2-Sp and its target genes in malarial sporozoites. Mol. Microbiol. 2010; 75(4):854. [PubMed: 20025671]
11. Iwanaga S, et al. Identification of an AP2-family Protein That Is Critical for Malaria Liver Stage Development. PLoS. ONE. 2012; 7(11):e47557. [PubMed: 23144823]
12. Flueck C, et al. Plasmodium falciparum heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. PLoS. Pathog. 2009; 5(9):e1000569. [PubMed: 19730695]
13. Burrill DR, Silver PA. Synthetic circuit identifies subpopulations with sustained memory of DNA damage. Genes Dev. 2011; 25(5):434. [PubMed: 21363961]
14. Shiels BR. Should I stay or should I go now? A stochastic model of stage differentiation in Theileria annulata. Parasitol Today. 1999; 15(6):241–245. [PubMed: 10366832]
15. Lopez-Rubio JJ, Mancio-Silva L, Scherf A. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. Cell Host. Microbe. 2009; 5(2):179. [PubMed: 19218088]
16. Heo JB, Sung S. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science. 2011; 331:76–9. [PubMed: 21127216]
17. Cameron, et al. Plasticity in transmission strategies of the malaria parasite, Plasmodium chabaudi: environmental and genetic effects. Evol Appl. Feb; 2013 6(2):365–76. [PubMed: 23467678]
18. Ikadai H, et al. Transposon mutagenesis identifies genes essential for Plasmodium falciparum gametocytogenesis. Proc. Natl. Acad. Sci. U. S. A. 2013
19. Janse CJ, et al. Variation in karyotype and gametocyte production during asexual multiplication of Plasmodium berghei. Acta Leiden. 1991; 60(1):43. [PubMed: 1820712]
20. Bruce MC, Alano P, Duthie S, Carter R. Commitment of the malaria parasite Plasmodium falciparum to sexual and asexual development. Parasitology. 1990; 100:191–200. [PubMed: 2189114]
21. Pfander C, et al. A scalable pipeline for highly effective genetic modification of a malaria parasite. Nat Methods. 2011; 8:1078–82. [PubMed: 22020067]
22. Kafsack BF, Painter HJ, Llinás M. New Agilent platform DNA microarrays for transcriptome analysis of Plasmodium falciparum and Plasmodium berghei for the malaria research community. Malar J. 2012; 11:187. [PubMed: 22681930]
23. Dearsly AL, Sinden RE, Self IA. Sexual development in malarial parasites: gametocyte production, fertility and infectivity to the mosquito vector. Parasitology. 1990; 100:359–368. [PubMed: 2194152]
24. Franke-Fayard B, Trueman H, Ramesar J, Mendoza J, van der Keur M, et al. A Plasmodium berghei reference line that constitutively expresses GFP at a high level throughout the complete life cycle. Mol Biochem Parasitol. 2004; 137:23–33. [PubMed: 15279948]
25. Khan SM, Franke-Fayard B, Mair GR, Lasonder E, Janse CJ, et al. Proteome analysis of separated male and female gametocytes reveals novel sex specific Plasmodium biology. Cell. 2005; 121:675–687. [PubMed: 15935755]
26. van Dijk MR, Janse CJ, Thompson J, Waters AP, Braks JA, et al. A central role for P48/45 in malaria parasite male gamete fertility. Cell. 2001; 104:153–164. [PubMed: 11163248]

27. Mair GR, et al. Universal features of post-transcriptional gene regulation are critical for Plasmodium zygote development. PLoS. Pathog. 2010; 6(2):e1000767. [PubMed: 20169188]
28. Ponzi M, et al. Egress of Plasmodium berghei gametes from their host erythrocyte is mediated by the MDV-1/PEG3 protein. Cell Microbiol. 2009; 11(8):1272. [PubMed: 19438517]
29. Orr RY, Philip N, Waters AP. Improved negative selection protocol for Plasmodium berghei in the rodent malarial model. Malaria J. 2012; 11:103.
30. Janse CJ, Ramesar J, Waters AP. High-efficiency transfection and drug selection of genetically transformed blood stages of the rodent malaria parasite Plasmodium berghei. Nat Protoc. 2006; 1:346–356. [PubMed: 17406255]
31. Sinden, R. Molecular Biology of Insect Diseases Vectors: A Methods Manual. Crampton, JM.; Beard, C. Ben; Louis, C., editors. Chapman and Hall; London: 1997. p. 67-91.
32. Quail MA, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent. BMC Genomics. 2012; 13:341. [PubMed: 22827831]
33. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–829. [PubMed: 18349386]
34. Swain MT, et al. A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes. Nat Protoc. 2012; 7:1260–84. [PubMed: 22678431]
35. Assefa S, et al. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics. 2009; 25:1968–1969. [PubMed: 19497936]
36. Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biology. 2010; 11:R41. [PubMed: 20388197]
37. Otto TD, et al. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. Bioinformatics. 2010; 26(14):1704–1707. [PubMed: 20562415]
38. Otto TD, Dillon GP, et al. RATT: Rapid Annotation Transfer Tool. Nucleic Acids Res. 2011; 39:e57. [PubMed: 21306991]
39. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16): 2078–9. [PubMed: 19505943]
40. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–303. [PubMed: 20644199]
41. Campbell TL, et al. Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. PLoS. Pathog. 2010; 6:e1001165. [PubMed: 21060817]
42. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW 3rd, et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol. 2006; 24:1429–1435. [PubMed: 16998473]
43. Berger MF, Bulyk ML. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc. 2009; 4:393–411. [PubMed: 19265799]
44. Workman CT, Yin Y, Corcoran DL, Ideker T, Stormo GD, et al. enoLOGOS: a versatile web tool for energy normalized sequence logos. Nucleic Acids Res. 2005; 33:W389–392. [PubMed: 15980495]
45. Pfander C, Anar B, Brochet M, Rayner JC, Billker O. Recombination-mediated genetic engineering of Plasmodium berghei DNA. Methods Mol Biol. 2013; 923:127–38. [PubMed: 22990774]
46. Zhang Y, Buchholz F, Muyrers JP, Stewart AF. A new logic for DNA engineering using recombination in Escherichia coli. Nat Genet. 1998; 20:123–128. [PubMed: 9771703]
47. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2012. ISBN 3-900051-07-0. URL http://www.R-project.org/
48. Smyth, GK. Limma: linear models for microarray data. In: Gentleman, R.; Carey, V.; Dudoit, S.; Irizarry, R.; Huber, W., editors. Bioinformatics and Computational Biology Solutions using R and Bioconductor. Springer, New York: 2005. p. 397-420.

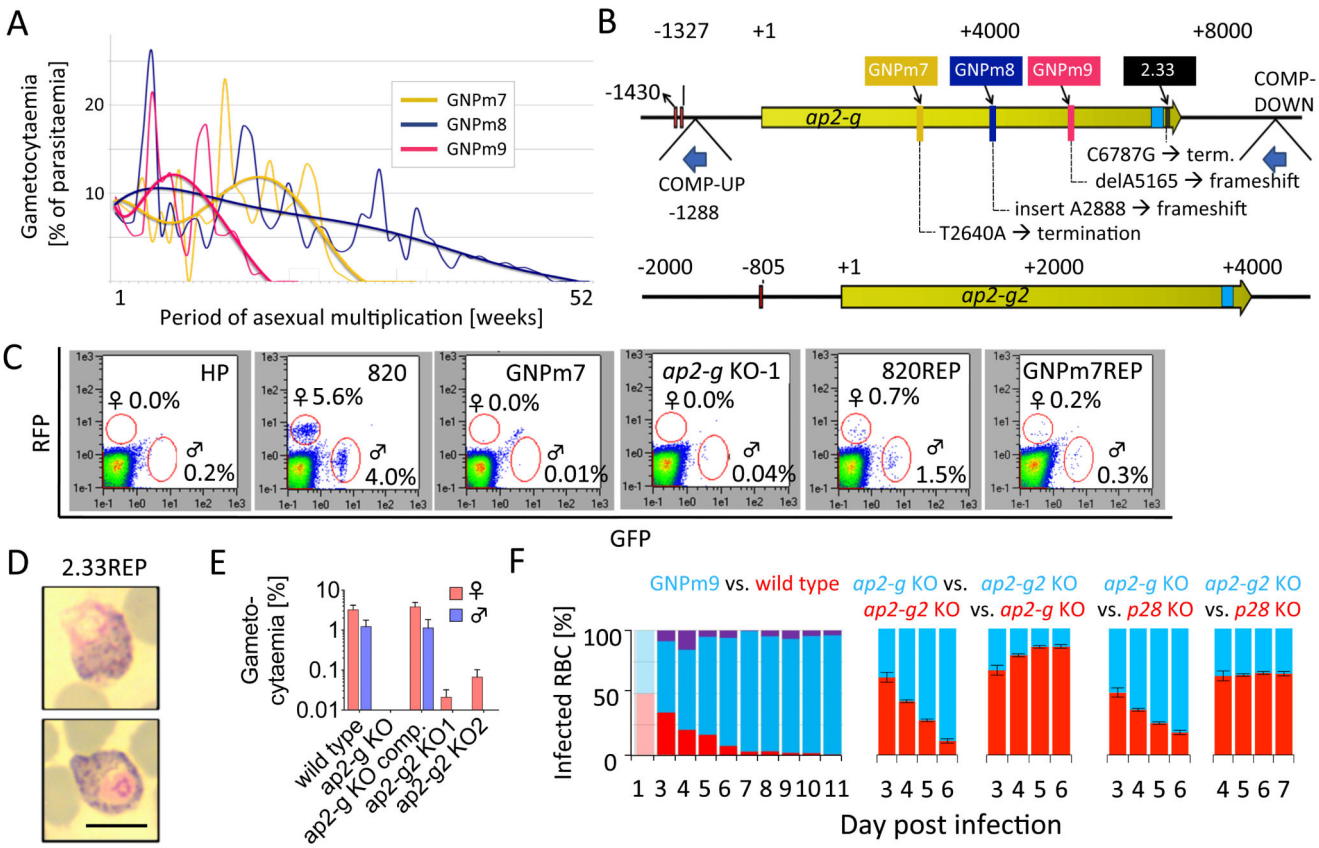


**Figure 1. Identification of mutations in PBANKA_143750 that account for the repeated spontaneous loss of commitment to gametocytogenesis**

A. Gametocyte production during a year of continuous mechanical passage of *P. berghei*. Best-fit polynomial trend (thick) lines of gametocytaemia on individual weekly observations (thin lines).

B. Open reading frames (ORF, in yellow) of *Pbap2-g* (PBANKA_143750) and *Pbap2-g2* (PBANKA_103430) with point mutations in new GNP lines from panel A and the long-established line 2.33. Predicted DNA Binding Domains (DBD, light blue) and DBD recognition motifs for PBAP2-G upstream of each ORF (brown bars) are indicated. Dark blue arrows show integration sites for selectable marker cassettes as used for genetic complementation of GNPs (COMP-DOWN) or to disrupt the promoter (COMP-UP). Numbering is relative to position 1 of the ORF.

C. FACS analyses of male and female gametocyte numbers (red circled areas) expressed as a percentage of the total parasitized cell counts. From left: *P. berghei* ANKA HP line (which lacks GFP/RFP reporters thus having no fluorescent signal and from which all subsequent lines reported in this study were derived) served as a negative control. Line 820 is the reporter line from which GNP mutants and a targeted KO (using vector PbGEM-072446) were derived. 820REP and GNPm7REP were generated with the COMP-DOWN complementation vector.

D. Giemsa-stained gametocytes in GNP line 2.33 (G756) repaired by the COMP-DOWN construct and after a single transmission through mosquitoes. Scale bar is 6 μm.




Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

E. Gametocyte quantification from manual counting in Giemsa-stained blood smears of an independently produced *pbap2-g* deletion mutant before and after complementation with the DS vector and of two independent *pbap2-g2* ko mutants. Error bars show standard deviations from 3 replicates. The loss of gametocytes from the KO mutants was significant ($p < 0.05$).

F. Relative growth kinetics of GNPm9, *pbap2-g*$^{-}$ and *pbap2-g2*$^{-}$ lines determined by flow cytometry. i) Cloned GNPm9 constitutively expressing CFP (line GNPm9-CFP) was mixed in a 1:1 ratio with wild type (PBANKA HP) producer line constitutively expressing RFP (line WT-RFP). The daily % of the population expressing either RFP (red), CFP (blue) or both (purple) reflecting mixed-multiply infected cells was calculated. ii) Deletion vectors for *ap2-g, ap2-g2* or *p28* (control gene for neutral growth rate) were transfected in GFP or mCherry expressing lines (blue and red bars, respectively) and the relative abundance of each mutant determined in mixed infections of uncloned parasites. Error bars show standard deviations from 3 biological replicates. The competitive advantage was significant for the *ap2-g*$^{-}$ ($p < 0.01$) but not *ap2-g2*$^{-}$ parasites (two tailed Student's T-test for change in relative abundance).

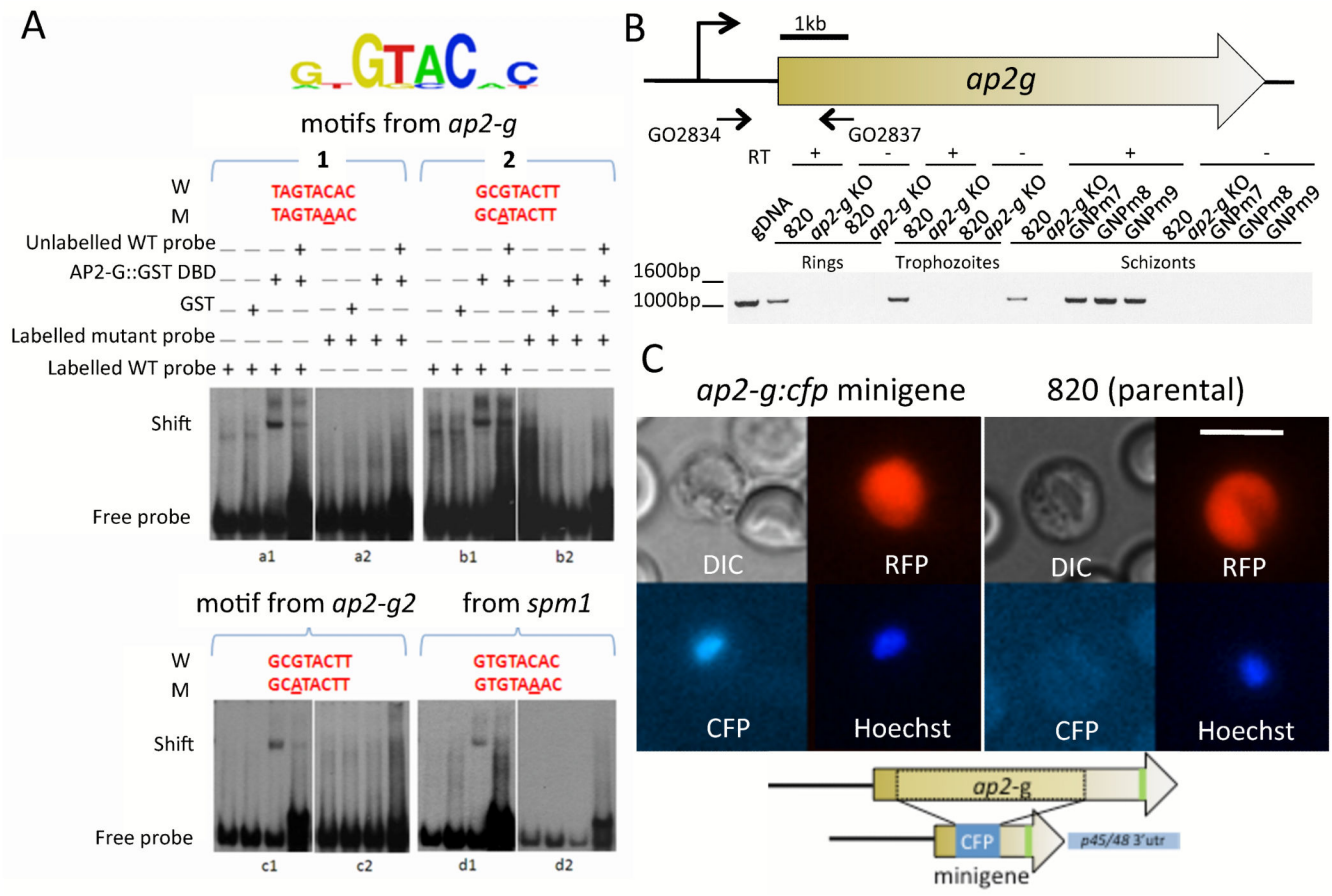


**Figure 2. Characterisation of the DNA binding specificity, expression and subcellular localisation of PBAP2-G**

A. Top. Protein Binding Microarray (PBM) determination of the DNA binding recognition preference of recombinant DNA Binding Domain (DBD) of PBAP2-G. Bottom. Electrophoretic Mobility Shift Assay (EMSA) in which a shift indicates whether AP2-G DBD binds to double stranded DNA containing wild type (W) or mutated (M) motifs (panels a1-d1and a2-d2, respectively) from the upstream regions of *Pbap2-g* itself, *Pbap2-g2*, and position −610 of a hypothetical gene *spm1* (Sub-Pellicular Microtubule protein-1, PBANKA_081070).

B. Expression analysis by RT-PCR of *ap2-g* in targeted and spontaneous *ap2-g* mutants and the wild type control line, 820. The 1.15 kb product indicates lack of transcript only in the targeted KO line. RT = reverse transcriptase. Primer positions were as shown in the schematic. See Fig. S7 for *ap2-g* transgene expression data. (N=3)

C. Localisation of *pbap2-g* mini gene-product to the nucleus of *P. berghei* female gametocytes. CFP was sandwiched between the N-terminal 300 bp and at the C-terminal 800 bp of *pbap2-g* including the DBD and expressed from 2 kb of the *pbap2-g* promoter in line 820. Expression was only detected in the nuclei of female gametocytes (>50 observations in 3 experiments). It is the C terminal segment that determines the nuclear localisation of PBAP2-G (Fig. S8). Scale bar = 6 μm. Cartoon is not to scale.




Europe PMC Funders Author Manuscripts

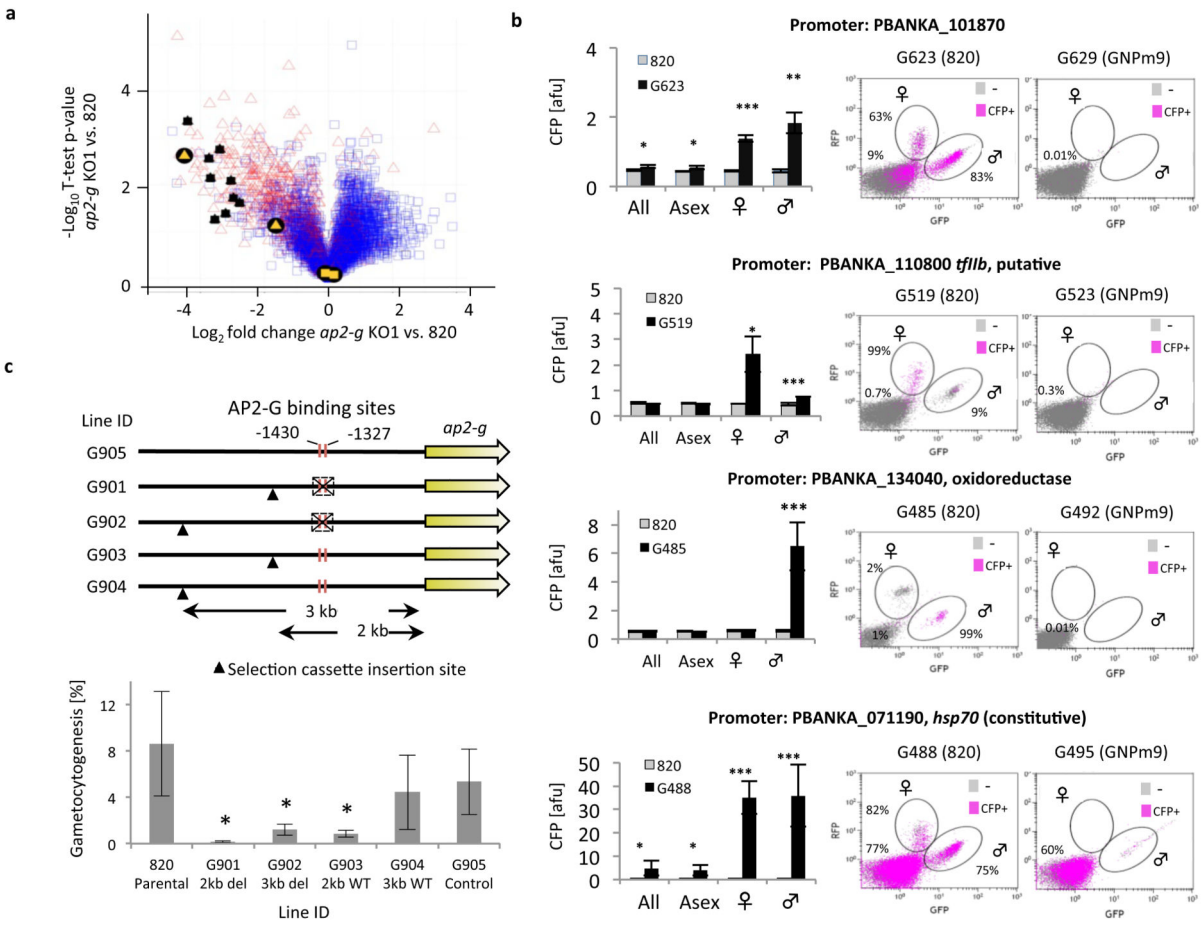


**Figure 3. *Pbap2-g* acts upstream of gametocyte gene transcription**

A. Volcano plot of $\log_2$ fold change in gene expression in schizonts of *ap2-g* KO1 (whole ORF deletion) vs. WT line 820 against significance of change (−log10 t-test). Red triangles indicate genes upregulated in gametocytes compared to schizonts. Black and yellow shapes are genes detailed in figure 3A and 3C, respectively.

B. Reporter gene expression constructs were transfected into the GNPm9 and 820 control clones to confirm gametocyte gene specific promoters. Reporters contained 2 kb of upstream sequence from the indicated genes driving CFP expression with a constitutive 3′ UTR. Bar plots show CFP measured by flow cytometry over three days in the 820 line. Life cycle stages (asexual, male and female) are separated based on GFP or RFP expression. Mean of three measurements (geometric mean CFP fluorescence) +/− SD, *P<0.05, ** P<0.01 ***P<0.001 by 2-tailed t-test. Flow cytometry plots are shown for CFP expression of reporters in 820 (parental) (left) or GNPm9 (right) lines. Plots show GFP (X-axis) vs. RFP (Y axis) expression for all infected red blood cells and CFP expression in magenta. Numbers on each plot represent the percentage of events within each gate that are positive for CFP (see also Figure S8).

C. Deletion studies in the *pbap2-g* promoter provide support for a role of PBAP2-G binding motifs in the positive feedback regulation of *pbap2-g* expression. Top: DNA constructs containing a selectable marker were integrated into the promoter region of *ap2-g* in




Europe PMC Funders Author Manuscripts

PBANKA 820. The constructs either deleted 207 bp surrounding the two instances of the AP2-G binding motif at the positions indicated (G901 and G902) or did not (G903 and G904). Two sites of selectable marker integration were tested, 2 and 3 kbp upstream of the ORF of *ap2-g*. Additionally interruption at −1288 upstream of the ORF of *ap2-g* was shown to disrupt gametocytogenesis (Fig. S4F). Control line G905 was transfected with a reporter construct targeted to the 230p locus and known not to affect gametocytogenesis. Bottom: Gametocytaemia was measured on consecutive days by flow cytometry once the parasitaemia reached >1%. Mean +/− SD shown, * $P<0.05$ compared to 820 parental (2-tailed t-Test). Data shown are pooled from three days' observations and representative of three independent experiments.

Europe PMC Funders Author Manuscripts

## Table 1

### Changes in gene expression in mutants

Gene expression was determined on Agilent microarrays for *invitro* cultured schizonts, comparing pooled GNP clones and targeted mutants to their parental control lines. $Log_2$ fold changes are shown for the top 10 genes with good functional annotation that were most strongly deregulated in the targeted mutant *ap2-g* KO1

| Gene ID | Description | Rank | GNP | *p* | *ap2-g* KO2 | *ap2-g2* KO1 |
|---|---|---|---|---|---|---|
| 051500 | 25 kDa ookinete surface antigen | 1 | −4.56 | *2.5E-02* | −4.88 | −1.72 |
| 051490 | 28 kDa ookinete surface antigen | 2 | −3.48 | *2.9E-02* | −6.28 | −2.37 |
| 133370 | phosphodiesterase delta | 125 | −3.61 | *1.3E-02* | −3.89 | −1.32 |
| 121910 | heat shock protein 90 | 175 | −3.34 | *7.6E-02* | −3.67 | −1.93 |
| 142170 | secreted ookinete protein, putative | 62 | −3.95 | *1.0E-01* | −3.98 | −1.42 |
| 131950 | LCCL domain-containing protein CCP2 | 64 | −3.09 | *6.2E-02* | −3.79 | −1.31 |
| 146300 | osmiophilic body protein | 232 | −1.63 | *1.2E-01* | −2.60 | −0.27 |
| 112040 | Pfs77 homologue, putative | 52 | −2.68 | *3.4E-02* | −3.50 | −0.78 |
| 134040 | oxidoreductase, putative | 327 | −4.59 | *5.9E-02* | −2.80 | −1.77 |
| 123130 | metabolite/drug transporter, putative | 26 | −3.31 | *5.0E-02* | −2.82 | −1.46 |

Gene IDs are given without their PBANKA_ prefix. Rank refers to the absolute expression rank among 4553 genes in purified gametocytes determined from three biological replicates. Expression data are means from three biological replicates for each mutant. *p* denotes the p-value adjusted for multiple testing. For the complete data and all p values see Table S6.