LETTER

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 Llinás[3,4]†, Matthew Berriman[2], Oliver Billker[2] & Andrew P. Waters[1]

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 apicomplexan AP2 (ApiAP2) family of DNA-binding proteins, is essential for the commitment of asexually replicating forms to sexual development in *Plasmodium 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, here 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 with a positive feedback loop involving PbAP2-G that could 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]. In a parasite line that produces fluorescently tagged gametocytes[2,3] 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, Supplementary Fig. 1 and Supplementary Table 1).

Subsequent developmental stages (gametes, ookinetes) were absent and none of the GNP lines could be transmitted through mosquitoes (Supplementary Fig. 2 and Supplementary Table 2). Whole-genome sequencing of these and an existing GNP line (ANKA 2.33) revealed numerous single nucleotide polymorphisms (SNPs) and insertions or deletions (indels) per line (Supplementary Fig. 3 and Supplementary Table 3); however, only a single gene, PBANKA_143750, carried a different and therefore independent nonsense or missense mutation in each line (Fig. 1b). PBANKA_143750 (here termed *pbap2-g*) encodes a putative transcription factor predicted to be composed of 2,330 amino acids with a single 55-amino-acid AP2 class DNA-binding domain (DBD) at its carboxy terminus (Fig. 1b). PbAP2-G belongs to the 27-strong[4,5] *Plasmodium* ApiAP2 family of transcription factors, themselves part of the larger Apetala 2/ethylene response factor (AP2/ERF) family of transcription factors 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 wild-type copy (generating GNPm7REP, GNPm8REP, GNPm9REP and 2.33REP) or genetic complementation of a targeted deletion mutant of *pbap2-g* (Fig. 1c and Supplementary Fig. 4a–g). Functionality of the restored gametocytes was demonstrated in GNPm7REP and 2.33REP by transmission through mosquitoes (Fig. 1d and Supplementary Table 4). 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 wild-type *P. berghei* and *pbap2-g2*$^{-}$ parasites, which had wild-type growth rates (Fig. 1e and Supplementary Fig. 5). *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, whereas 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] recognized closely related DNA motifs (Fig. 2a and Supplementary Table 5) identical to the previously derived motif for the DBD from the orthologous ApiAP2 protein of *Plasmodium falciparum* (PF3D7_1222600)[6], confirming that both DBDs bind primarily to the same (GxGTACxC) motif (in which x denotes any residue). Electrophoretic mobility shift assay (EMSA) analyses (Fig. 2a) refined the motif to two 6-mers (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 2 kilobases (kb) upstream of 49% of all genes (2,359 of 4,803 considered), yet more frequently in genes designated as upregulated in gametocytes (246 (54%) of 452 genes; $P < 0.002$, hypergeometric test). 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 recognized 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 *pbap2-g* produced no detectable protein (Supplementary Fig. 6) yet gametocytogenesis was unaltered, implying that tagged PbAP2-G activity is unaffected. However, a truncated cyan fluorescent protein (CFP)-tagged transgene product could be detected in nuclei of female gametocytes (Fig. 2c and Supplementary Fig. 7).

Comparative microarray analyses showed that gametocyte-specific genes were highly enriched among the 500 most downregulated genes

[1] Wellcome Trust Centre for Molecular Parasitology, University of Glasgow, Glasgow G12 8QQ, UK. [2] Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [3] Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544-1014, USA. [4] Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA. †Present address: Department of Biochemistry and Molecular Biology and Center for Infectious Disease Dynamics, The Pennsylvania State University, State College, Pennsylvania 16802, USA.
*These authors contributed equally to this work.

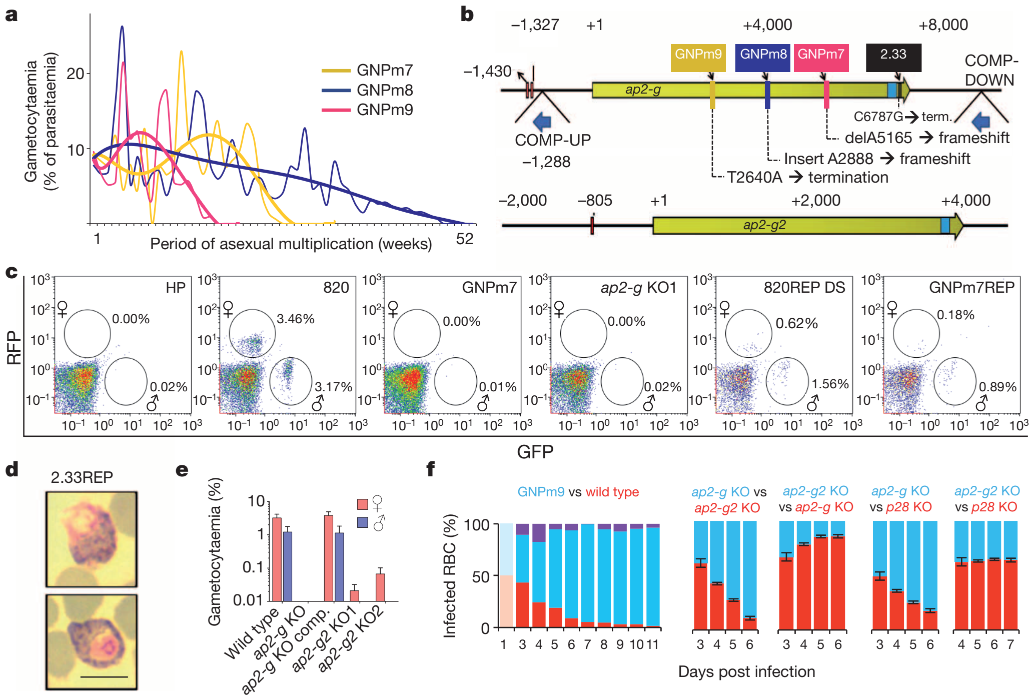

Figure 1 | Identification of mutations in *pbap2-g* that account for the repeated spontaneous loss of commitment to gametocytogenesis. **a**, Gametocyte production during a year of continuous mechanical passage of *P. berghei.* Best-fit polynomial trend (thick) lines of gametocytaemia on individual weekly observations (thin lines). **b**, Open reading frames (ORF) (yellow) of *pbap2-g* (PBANKA_143750) and *pbap2-g2* (PBANKA_103430) with point mutations in new GNP lines shown in **a** and the long-established line 2.33. Predicted DBDs (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 (circled areas) expressed as a percentage of the total parasitized cell counts. From left, *P. berghei* ANKA HP line (which lacks green (GFP) or red (RFP) fluorescent protein 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 knockout (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, 6 µm. **e**, Gametocyte quantification from manual counting in Giemsa-stained blood smears of an independently produced *pbap2-g* deletion mutant before and after complementation (comp.) with the DS (downstream) vector and of two independent *pbap2-g2* knockout mutants. Error bars show standard deviations from three replicates. The loss of gametocytes from the knockout mutants was significant ($P < 0.05$). **f**, Relative growth kinetics of GNPm9, *pbap2-g*$^{-}$ and *pbap2-g2*$^{-}$ lines determined by flow cytometry. Left, 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 percentage of the population expressing either RFP (red), CFP (blue) or both (purple; reflecting cells infected with multiple parasites) was calculated. Right four panels, deletion vectors for *pbap2-g*, *pbap2-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 three biological replicates. The competitive advantage was significant for the *pbap2-g*$^{-}$ ($P < 0.01$) but not the *pbap2-g2*$^{-}$ parasites (two tailed Student's *t*-test for change in relative abundance). RBC, red blood cell.

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 (Table 1 and Supplementary Fig. 6). Comparison of the transcriptomes of wild-type 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 downregulated (>2 s.d. reduced from the mean, Supplementary Table 6) in schizonts.

The activity of 18 promoters consistently downregulated in GNP lines, and which contain one or more candidate PbAP2-G-binding motifs, was analysed in wild-type and GNPm9 parasite backgrounds. Male, female or sex-specific genes downstream of AP2-G in the gametocyte developmental pathway were identified (Fig. 3b, Supplementary Fig. 8 and Supplementary Table 8). 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 (Supplementary Fig. 9). 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 *pbap2-g* 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 transcription factors in apicomplexan genomes, otherwise thought to be remarkably lacking in genes encoding transcription factors[8]. The majority of ApiAP2 transcription factors 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 have been demonstrated, namely for the ookinete (PbAP2-O[9]), sporozoite (PbAP2-S[10]) and liver stages (PbAP2-L[11]) of development. ApiAP2 proteins may also silence genes, possibly through maintenance of heterochromatin[12]. The AP2/ERF family members in *Plasmodium* are predicted to act singly or in combinations that control the continuation of the transcriptional

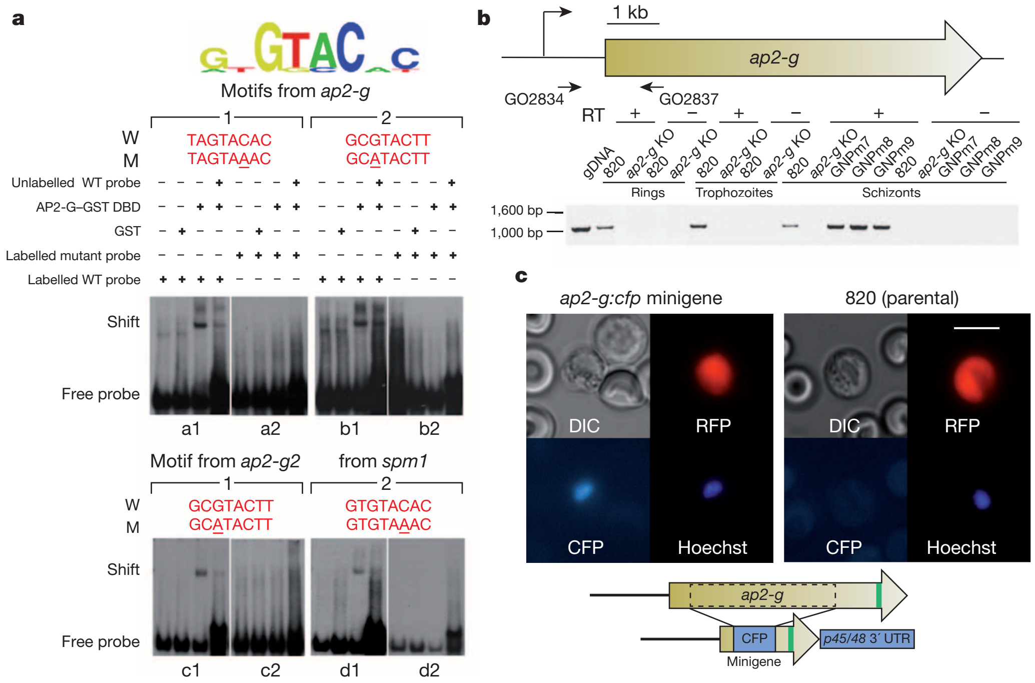

Figure 2 | Characterization of the DNA-binding specificity, expression and subcellular localization of PbAP2-G. a, Top, protein binding microarray determination of the DNA binding recognition preference of the recombinant DBD of PbAP2-G. GST, glutathione *S*-transferase. Bottom, EMSA in which a shift indicates whether the PbAP2-G DBD binds to double-stranded DNA containing wild-type (W) or mutated (M) motifs (panels a1–d1 and a2–d2, respectively) from the upstream regions of *pbap2-g* itself, *pbap2-g2*, and position −610 of the hypothetical gene *spm1* (subpellicular microtubule protein 1, PBANKA_081070). **b**, Expression analysis by reverse-transcriptase (RT)–PCR of *pbap2-g* in targeted and spontaneous *pbap2-g* mutants and the wild-type control line, 820. The 1.15-kb product indicates lack of transcript only in the targeted knockout line. Primer positions were as shown in the schematic. See Supplementary Fig. 7 for *pbap2-g* transgene expression data. $n = 3$. gDNA, genomic DNA. GO, Glasgow oligo. **c**, Localization of the *pbap2-g* minigene product to the nucleus of *P. berghei* female gametocytes. CFP was sandwiched between the N-terminal 300 base pairs (bp) and 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 three experiments). It is the C-terminal segment that determines the nuclear localization of PbAP2-G (Supplementary Fig. 8). Scale bar, 6 μm. Cartoon is not to scale. DIC, differential interference contrast.

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 (for example, gametocytogenesis) is probabilistic, its frequency being defined by the likelihood of the interaction of a fate-determining transcription factor 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 transcription factor 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 PfAP2-G production, could provide flexible control of gametocytogenesis in a manner that would also be amenable to environmental sensing[16,17]. AP2-G is, at present, unique within the apiAP2 transcription factor family in that it directs a change in developmental fate rather than merely progressing a lineage (Supplementary Fig. 10), 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*[19], even though models for the timing of commitment in the two parasites differ[20,21]. 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 (Supplementary Fig. 11). Thus these data identify the earliest known event in parasite transmission. Because it occurs in the blood of the host it is amenable to and suggests novel control strategies largely through drug development and

Table 1 | Changes in gene expression in mutants

Gene ID	Description	Rank	GNP	*P*	*pbap2-g* KO2	*pbap2-g2* KO1
051500	25-kDa ookinete surface antigen	1	−4.56	2.5×10^{-2}	−4.88	−1.72
051490	28-kDa ookinete surface antigen	2	−3.48	2.9×10^{-2}	−6.28	−2.37
133370	Phosphodiesterase delta	125	−3.61	1.3×10^{-2}	−3.89	−1.32
121910	Heat-shock protein 90	175	−3.34	7.6×10^{-2}	−3.67	−1.93
142170	Secreted ookinete protein, putative	62	−3.95	1.0×10^{-1}	−3.98	−1.42
131950	LCCL domain-containing protein CCP2	64	−3.09	6.2×10^{-2}	−3.79	−1.31
146300	Osmiophilic body protein	232	−1.63	1.2×10^{-1}	−2.60	−0.27
112040	Pfs77 homologue, putative	52	−2.68	3.4×10^{-2}	−3.50	−0.78
134040	Oxidoreductase, putative	327	−4.59	5.9×10^{-2}	−2.80	−1.77
123130	Metabolite/drug transporter, putative	26	−3.31	5.0×10^{-2}	−2.82	−1.46

Gene expression was determined on Agilent microarrays for *in vitro*-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 *pbap2-g* KO1. Gene IDs are given without their PBANKA_ prefix. Rank refers to the absolute expression rank among 4,553 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 Supplementary Table 6.

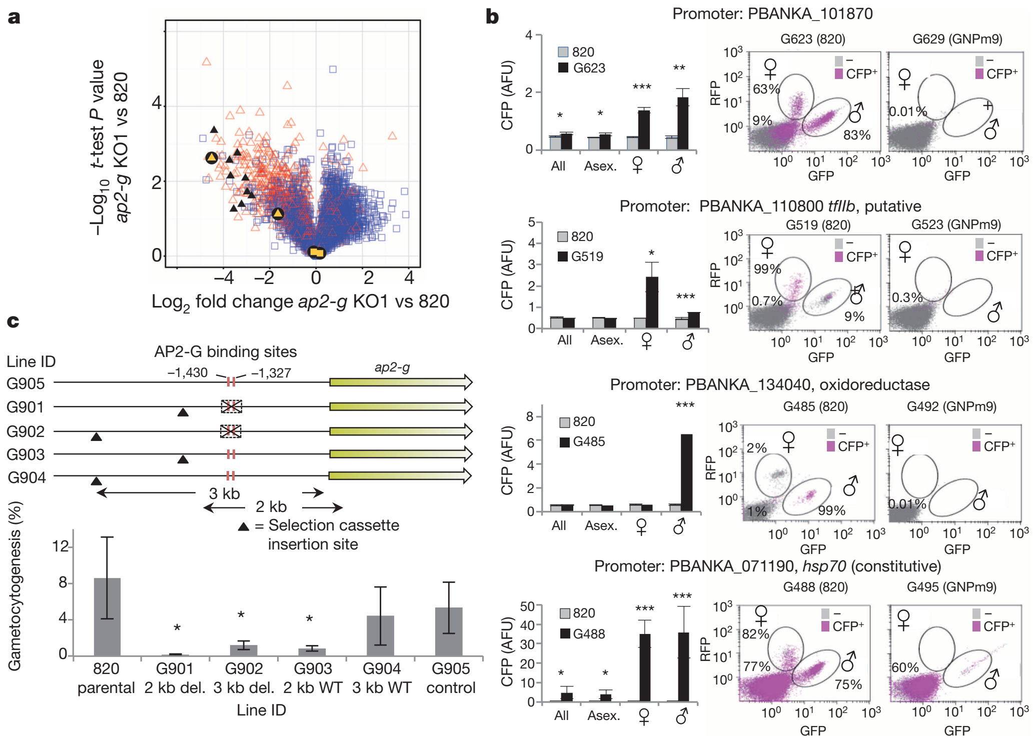

Figure 3 | *pbap2-g* acts upstream of gametocyte gene transcription. **a**, Volcano plot of $\log_2$ fold change in gene expression in schizonts of *pbap2-g* KO1 (whole ORF deletion) versus wild-type line 820 against significance of change (−log10 *t*-test). Red triangles indicate genes upregulated in gametocytes compared to schizonts. Black and yellow shapes are genes detailed in Table 1 and Fig. 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′ untranslated region. Bar plots show CFP measured by flow cytometry over 3 days in the 820 line. Life cycle stages (asexual, male and female) are separated on the basis of GFP or RFP expression. Mean of three measurements (geometric mean CFP fluorescence) ± s.d.; $*P < 0.05$, $** P < 0.01$, $***P < 0.001$, two-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) versus 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 Supplementary Fig. 8). **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 *pbap2-g* in PBANKA 820. The constructs either deleted 207 bp surrounding the two instances of the PbAP2-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 kb upstream of the ORF of *pbap2-g*. In addition, interruption at −1,288 upstream of the ORF of *pbap2-g* was shown to disrupt gametocytogenesis (Supplementary Fig. 4f). Control line G905 was transfected with a reporter construct targeted to the *p230p* locus and known not to affect gametocytogenesis. Bottom, gametocytaemia was measured on consecutive days by flow cytometry once the parasitaemia reached >1%. Mean ± s.d. shown, $*P < 0.05$ compared to 820 parental (two-tailed *t*-test). Data shown are pooled from 3 days' observations and representative of three independent experiments.

offers some strategic value in the prevention of sexual development and reduction of transmission.

METHODS SUMMARY

P. berghei ANKA parasites were maintained in female Theiler's original (TO) mice (6–8 weeks old) under appropriate Home Office licences. A fluorescent reporter line 820 (ref. 3) for male (green) and female (red) gametocytes was transmitted weekly by blood passage into a new host for up to 52 weeks in 10 parallel lines and gametocytaemia assessed weekly by flow cytometry. Whole-genome sequencing was followed by *de novo* assembly and variant calling. Targeted gene knockouts were generated using traditional plasmids or *Plasmo*GEM vectors[22]. GNP phenotypes were confirmed by a variety of methods. Genetic complementation was by ends-out recombination over the region mutated in GNP clones and confirmed functionally by FACS and mosquito passage. A *pbap2-g* DBD–GST fusion protein was used in protein binding microarray analysis as described[6]. The purified GST-recombinant protein was used in EMSA assays with 60-mer biotinylated annealed oligonucleotides. Microarray analysis was performed on total RNA on an Agilent array[23] and data submitted to the Gene Expression Omnibus (GEO) database. Reporter constructs were transfected into 820 and GNPm9. Reporter expression was monitored by FACS over several days. The promoter of *pbap2-g* was modified by ends-out integration into 820 and gametocytaemia monitored over several days using flow cytometry.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 April; accepted 19 December 2013.
Published online 23 February 2014.

1. Janse, C. J. *et al. Plasmodium berghei*: *in vivo* generation and selection of karyotype mutants and non-gametocyte producer mutants. *Exp. Parasitol.* **74,** 1–10 (1992).
2. Mair, G. R. *et al.* Universal features of post-transcriptional gene regulation are critical for *Plasmodium* zygote development. *PLoS Pathog.* **6,** e1000767 (2010).
3. Ponzi, M. *et al.* Egress of *Plasmodium berghei* gametes from their host erythrocyte is mediated by the MDV-1/PEG3 protein. *Cell. Microbiol.* **11,** 1272–1288 (2009).
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.* **33,** 3994–4006 (2005).
5. Painter, H. J., Campbell, T. L. & Llinas, M. The Apicomplexan AP2 family: integral factors regulating *Plasmodium* development. *Mol. Biochem. Parasitol.* **176,** 1–7 (2011).
6. Campbell, T. L. *et al.* Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathog.* **6,** e1001165 (2010).
7. Berger, M. F. & Bulyk, M. L. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature Protocols* **4,** 393–411 (2009).
8. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).
9. Yuda, M. *et al.* Identification of a transcription factor in the mosquito-invasive stage of malaria parasites. *Mol. Microbiol.* **71,** 1402–1414 (2009).
10. Yuda, M. *et al.* Transcription factor AP2-Sp and its target genes in malarial sporozoites. *Mol. Microbiol.* **75,** 854–863 (2010).
11. Iwanaga, S. *et al.* Identification of an AP2-family protein that is critical for malaria liver stage development. *PLoS ONE* **7,** e47557 (2012).

12. Flueck, C. *et al.* *Plasmodium falciparum* heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. *PLoS Pathog.* **5,** e1000569 (2009).
13. Burrill, D. R. & Silver, P. A. Synthetic circuit identifies subpopulations with sustained memory of DNA damage. *Genes Dev.* **25,** 434–439 (2011).
14. Shiels, B. R. Should I stay or should I go now? A stochastic model of stage differentiation in *Theileria annulata*. *Parasitol. Today* **15,** 241–245 (1999).
15. Lopez-Rubio, J. J., 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* **5,** 179–190 (2009).
16. Heo, J. B. & Sung, S. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science* **331,** 76–79 (2011).
17. Cameron, A., Reece, S. E., Drew, D. R., Haydon, D. T. & Yates, A. J. Plasticity in transmission strategies of the malaria parasite, *Plasmodium chabaudi*: environmental and genetic effects. *Evol. Appl.* **6,** 365–376 (2013).
18. Ikadai, H. *et al.* Transposon mutagenesis identifies genes essential for *Plasmodium falciparum* gametocytogenesis. *Proc. Natl Acad. Sci. USA* **110,** E1676–E1684 (2013).
19. Kafsack, B. F. C. *et al.* A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature* http://dx.doi.org/10.1038/nature12920 (this issue).
20. Janse, C. J. *et al.* Variation in karyotype and gametocyte production during asexual multiplication of *Plasmodium berghei. Acta Leiden.* **60,** 43–48 (1991).
21. Bruce, M. C., Alano, P., Duthie, S. & Carter, R. Commitment of the malaria parasite *Plasmodium falciparum* to sexual and asexual development. *Parasitology* **100,** 191–200 (1990).
22. Pfander, C. *et al.* A scalable pipeline for highly effective genetic modification of a malaria parasite. *Nature Methods* **8,** 1078–1082 (2011).
23. Kafsack, B. F., Painter, H. J. & Llinás, M. New Agilent platform DNA microarrays for transcriptome analysis of *Plasmodium falciparum* and *Plasmodium berghei* for the malaria research community. *Malar. J.* **11,** 187 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. McBride for technical assistance with mosquito transmissions; A. R. Gomes, J. Tripathi, D. Vaughan and the III flow cytometry facility for assistance; C. Cairney and N. Keith at the Beatson Institute for use of their Agilent microarray scanner; and A. Cortes and D. Baker for reading of the manuscript. A.P.W. is funded by the Wellcome Trust (ref. 083811/Z/07/Z). The Wellcome Trust Centre for Molecular Parasitology is supported by core funding from the Wellcome Trust (085349). A.S. was a student of the University of Glasgow Wellcome Trust 4-year PhD Programme Molecular Functions in Disease. A.P.W., O.B. and M.B. are members of Evimalar (ref. 242095), which funds the work of T.D.O. Work at the Sanger Institute was funded by grants from the Wellcome Trust (098051) and the Medical Research Council (G0501670). M.L. is funded by the National Institutes of Health (R01 AI076276) and the Centre for Quantitative Biology (P50GM071508). B.F.C.K. was supported by a Howard Hughes Medical Institute fellowship of the Damon Runyon Cancer Research Foundation.

Author Contributions A.P.W. and O.B. directed the research. A.S. generated the GNP clones, performed some of the EMSA analyses, made *pbap2-g* gene knockout lines and complementation lines and analysed the latter. K.R.H. performed microarray analyses, generated reporter and minigene constructs, made transgenic parasites and analysed them, performed competition experiments. K.K.M. made the complementation construct, generated and analysed knockout and complemented lines for *pbap2-g* and *pbap2-g2* and performed and analysed competition and microarray experiments. C.P. generated knockout lines for *pbap2-g* and *pbap2-g2* and performed the initial parasitological analysis. E.B. generated recombinase engineered constructs for use at Wellcome Trust Sanger Institute and University of Glasgow. A.L.G. and A.A.R. performed expression analyses. N.J.D. performed statistical analyses of motif distribution and assisted with the microarray analyses. R.C. performed the complementation experiments and transmission experiments. A.E.W. performed EMSA analyses and generated constructs used in the analysis. T.D.O. and M.B. generated the GNP sequence data and SNP analyses. M.L. and B.F.C.K. performed microarray analyses, M.L. and A.E.W. performed EMSA analyses and generated recombinant PbAP2-G DBD. A.P.W., O.B., A.S., K.R.H. and K.K.M. wrote the paper.

Author Information Microarray data has been submitted to the GEO database under accession numbers GSE52859 and GSE53246. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to O.B. (OB4@sanger.ac.uk) or A.P.W. (Andy.Waters@glasgow.ac.uk).

METHODS

Parasite lines and methods. *P. berghei* ANKA HP was obtained from C. Janse at Leiden University Medical Centre and was originally referred to as clone 15Cy1A (Leiden Malaria Group website). Line 820 was generated from HP. *P. berghei* ANKA clone 2.33 is a non-gametocyte-producing clone line reported in 1990 and is now widely distributed and grown by mechanical passage[24]. All infections were performed on female Theiler's original (TO) mice (age 6–8 weeks; weight 25–30 g) according to Home Office licence regulations and the local ethical committees. All animals were assigned to experiments without pre-selection and no blind assignments were performed. Serial passage of freshly cloned *P. berghei* reference line 820 m1cl1 (ref. 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[25–29]. 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, GNPm8 and GNPm9 were cloned by limiting dilution, clones subjected to negative selection[30] to remove the selectable marker residual in the GFP:RFP selection cassette and cloned once more. Each parasite cloning procedure used 10 mice, and mice were infected by intravenous 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, GNPm8 and GNPm9.

Transfection of GFP- and RFP-expressing 'wild-type' parasites from the *P. berghei* line 820 with linearized targeting constructs, selection and cloning of the mutant parasites were performed according to procedures described previously[31]. Genotypic analysis of transfected parasites was performed by Southern analysis of chromosomes separated by field-inversion gel electrophoresis and using diagnostic PCR on genomic DNA. Details of the primers used for PCR are shown in Supplementary Table 9. Phenotype analysis of mutant parasites during blood-stage development, quantification of gametocyte production and ookinete development *in vitro* was performed using standard methods as described previously[26–29]. Mosquito-stage development was analysed in *Anopheles stephensi* mosquitoes using standard methods of mosquito infection, analysis of oocyst and sporozoite production and sporozoite infectivity to TO mice[32]. 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 until day 8 after infection. Infectivity was recorded as 'wild type' if mice developed a parasitaemia 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 Fig. 1d; similar results were seen on 3 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[33]. 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 Supplementary Table 3. 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[34] 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 average length (N50) of 240 kb. We processed the assembly as described in the post-assembly genome-improvement toolkit protocol[35]. In short, scaffolds were ordered with ABACAS[36] 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[37] was used to close 469 (61%) of the 774 sequencing gaps. Single-base and indel errors were corrected using ICORN[38]. This corrected 1,067 single-base errors and 92 indels. 1,589 positions had heterozygous calls, which represented collapsed repeats, mostly in *P. berghei* interspersed repeat (*bir*) genes. Last, the annotation of the *P. berghei* ANKA reference genome was transferred onto the improved *P. berghei* 820 assembly using RATT[39] (Assembly option). In total, 4,821 of the 4,938 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[40], variation was called with GATK[41] (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 pipeline for whole-genome sequencing and identification of single nucleotide polymorphisms is summarized in Supplementary Fig. 12.

Variant calling in *Plasmodium* from re-sequencing data are inherently noisy, owing 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 Supplementary Table 3 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. Neighbour joining tree was generated in CLC Genomics Workbench (version 6.5.1) using the Jukes-Cantor protein distance measure. Values shown are for 1,000 bootstrap iterations. The tree is rooted using the most distant *Arabidopsis thaliana* DBD Q9M0l0.2.

Recombinant protein production. N-terminal GST-fused extended ApiAP2 DBDs (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 Supplementary Table 10.

Protein binding microarray analysis. Protein binding microarray analyses were processed and analysed as described previously[42–45].

EMSAs. 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 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 synthesized and purchased from MWG Eurofins (Germany) as labelled (5′-biotinylated and 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 the DBD of AP2G. EMSAs were performed using the LightShift Chemiluminescent EMSA kit (Pierce). In brief, 2 μg of the purified GST fusion of PfAP2-G and PbAP2-G (in separate reactions) was pre-incubated with 0.02 pmol of the labelled probe in 20 μl of 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 (22 °C) for 10 min. 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 min 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 *pbap2-g* length-variable knockouts was performed to show successful integration of the selectable marker cassette at the desired genetic locus. In brief, approximately 10 μg of Plasmodipur (EuroProxima)-filtered and purified genomic DNA from lines 820 (wild type), G401cl1 (complete ORF knockout), G418cl6cl3 (DBD 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 h with NEB Buffer 4. For comparison with the wild-type line (820), gDNA from wild type and G401cl1, wild type and G418cl6cl3, and wild type and G529cl2 was double-digested with the High-Fidelity versions of NcoI and SpeI, NcoI and BamHI, and EcoRI and 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 and GU1059 for G401cl1, GU1416 and GU1417 for G418cl6cl3 and GU1414 and GU1415 for G529cl2. The membrane was washed three times 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 h.

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* messenger RNA (PBANKA_051490; 0.62 kb ORF) and normalized using *hsp70* mRNA probe (PBANKA_071190; 2.08 kb ORF), washed and exposed to a maximum resolution X-ray film (BioMax MR film; Kodak).

Recombineering methods. Gene knockout vectors for *pbap2-g* and *pbap2-g2* were submitted to the *Plasmo*GEM database as PbGEM-072446 and PbGEM-039238, respectively[22] where details of their construction can be found. Complementation vectors were made using the Red recombination system of phage *lambda* using published protocols[46]. First, *E. coli* harbouring *P. berghei* gDNA clone PbG01-2472c01, which carries a >11-kb genomic insert including *pbap2-g* in the pJAZZ-OK linear plasmid (Lucigen), were rendered competent for recombination by transfection with plasmid pSC101gbaA[47]. 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 Supplementary Table 11 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[46]. 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[30] and target regions for DXO integration into a *p230p* locus downstream of the GFP/RFP cassette in the 820 line[29]. Candidates for reporter analysis in the first batch (rep 1–14) were chosen on the basis of fold downregulation in GNP versus 820 schizont, the presence of at least one predicted AP2 binding motif (GTACxC or GxGTAC or GGTACxC) 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. 2 kb 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 2 kb fragment in a two-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 linearize 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 re-suspended 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 (4–12 µg per transfection) was mixed with Nycodenz-purified synchronous *P. berghei* schizont 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 h after transfection the parasites were placed on positive selection by including pyrimethamine (Sigma) in drinking water[31].

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 (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 re-suspended 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. 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 as forward scatter (FS) versus side scatter (SS) and gate E drawn to exclude debris. Events in gate E were plotted on FS versus FS (area) and gate J(1) drawn to exclude potentially autofluorescent doublets and clumps. Events in gate J(1) were plotted FS versus Ruby (DNA stain) and gate G drawn to select infected cells. Gate G was drawn on the basis of a negative (uninfected) control population stained in the same way and analysed on the same occasion (Supplementary Fig. 13a).

Events in gate G were plotted SS versus 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 versus RFP was plotted for all infected cells (events in G) and for only those falling into the CFP-positive gate. Gates drawn on female F (RFP-positive) and male M (GFP-positive) populations was used to calculate the percentage of each population that expresses CFP based on the number of cells in each gate in each plot.

For illustrative figures the infected population (G) was plotted on GFP versus RFP and those additionally falling into gate CFP-positive coloured magenta whereas those not CFP-positive were coloured grey. The percentage of the population within each gate expressing CFP (calculated as above) is indicated (Supplementary Fig. 13b).

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 visualized 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 *pbap2-g* promoter slightly downstream of two GxGTAC motifs led to a loss of gametocyte production. To investigate this further a series of constructs was made to target the *pbap2-g* endogenous promoter and mutate specifically in the region of these GxGTAC 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-cross-over homologous recombination method was used to create targeted interruptions of the *pbap2-g* endogenous promoter. The plasmid pL0035 was used, which contains a selection cassette including human *DHFR* driven by the *pbeef1aa* promoter surrounded by multiple cloning sites. Genomic fragments from the *pbap2-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 207-bp region containing the GxGTAC motifs was synthesized by MWG-Biotech with or without point mutations in the core motif. All regions are described by their distance from the *pbap2-g* gene start. A downstream integration fragment from bp-416 to bp-1,277 was cloned in using SmaI and EcoRI and an upstream integration region from bp-2,695 to bp-1,912 cloned in using HindIII and SacII. The region from bp-1,913 to bp-1,484 was cloned downstream of the selection cassette and in front of the downstream integration region using KpnI and EcoRV to create vector pG266 (2-kb deletion). Using SmaI and EcoRV the synthesized region from −1,913 to −1,484, either wild type or containing single point mutations in the G.GTAC motif, was cloned into vector pG266 to create pG298 (2-kb WT) or pG312 (2 kb MutA). Additionally a clone containing the wild-type 200-bp region in reverse orientation was selected pG299 (2-kb WT Rev). Subsequently the SmaI cloning site in pG298 was removed to created pG313 (2-kb WT-Sma). To extend the region of endogenous promoter remaining between the selection cassette and the *pbap2-g* gene an additional fragment from −2,870 to −1,913 was cloned into the KpnI site downstream of the selection cassette in pG266 and pG313 to create pG266+3 (3-kb del) and pG313+3 (3-kb WT). Constructs were linearized using HindIII and EcoRI, and approximately 10 µg of purified linear DNA was transfected in to *P. berghei* parasites (820 line) as described elsewhere.

Promoter interruption gametocytogenesis essays. 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. As the background gametocyte levels measurable using our methods in the parental 820 line varied from $\sim$3 to $\sim$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 ± s.d. 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 2-kb fragment immediately upstream of the *pbap2-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 300-bp fragment of *pbap2-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 *pbap2-g* incorporating the DBD was amplified from gDNA 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 *pbap2-g* promoter and first 300 bp 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 linearized 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 localization 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 red blood cells 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 intravenously into mice. Parasites were monitored daily by flow cytometry and after gating for infected cells the percentage 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 *pbap2-g* KO1, *pbap2-g2* KO and *p28* KO, the *Plasmo*GEM knockout vectors were transfected into the GFP- and mCherry-expressing parasites. Once the parasitaemia in transfected animals reached $\sim$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 (1×10^5 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 8×15k custom microarray (Agilent) providing coverage of the *P. berghei* genome at $>$1 probe per kb of coding sequence was used[23]. 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 collected at 24 + 6 h post infection and filtered through a magnetic column (variomacsD) to deplete of mature stages and gametocytes. For trophozoite-stage parasites, parasites were prepared as for ring stages were then cultured for a further 6 h. 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[23]. For GNP and *pbap2-g* KO1 a two-colour microarray hybridization was performed with a background pool of complementary DNA made from material from all life cycle stages (except late mosquito and liver stages). Parental control lines and experimental samples were then hybridized with the same background pool sample for all experiments. For *pbap2-g* KO2 and *pbap2-g* KO, the mutant samples were hybridized 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 (http//: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 *pbap2-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 *pbap2-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) versus 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 ($P < 2.2 \times 10^{-16}$, Kolmogorov–Smirnov test in R version 2.10), and the transcripts classed as down regulated in GNP lines were those 2 s.d. below the mean fold change. For plotting volcano plots (Fig. 3) a two-tailed *t*-test was performed on the independent replicates and a $-\log_{10}$ 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 *pbap2-g* KO2 and *pbap2-g2* the biological triplicates of each of the hybridizations (both mutants against the wild type and against each other) were processed using the R version 2.15.0 software[48] with limma package[49]. The data was background-corrected and normalized between the arrays (LOESS normalization). 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 correction.

For the gametocyte expression rank (Fig. 3a and Supplementary Table 6) the absolute intensity values from microarrays from three independent replicates of wild-type gametocytes was used and ranked from highest (1) to lowest ($\sim$4,553) expression rank. To test for the deregulation of the gametocyte-specific genes in all the strains, the enrichment in gametocyte-specific genes (expression rank 1 to 500) in the top 500 genes showing the highest fold change in each of the mutants was tested using the Fisher's exact test. Comparisons of the variances of the microarray data were carried out in R and all the variances were similar; none of the samples were significantly different ($P < 10 \times 10^{-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 2-kb region upstream of the transcription start site to the first base of the transcription start site (4,803 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 /GxGTAC|GTACxC/ and PbAP2-G2 was defined by orthology as /TGCxACC|GGTxGCA/; ref. 6) 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.

24. Dearsly, A. L., Sinden, R. E. & Self, I. A. Sexual development in malarial parasites: gametocyte production, fertility and infectivity to the mosquito vector. *Parasitology* **100,** 359–368 (1990).
25. Franke-Fayard, B. *et al.* A *Plasmodium berghei* reference line that constitutively expresses GFP at a high level throughout the complete life cycle. *Mol. Biochem. Parasitol.* **137,** 23–33 (2004).
26. Khan, S. M. *et al.* Proteome analysis of separated male and female gametocytes reveals novel sex specific *Plasmodium* biology. *Cell* **121,** 675–687 (2005).
27. van Dijk, M. R. *et al.* A central role for P48/45 in malaria parasite male gamete fertility. *Cell* **104,** 153–164 (2001).

28. Mair, G. R. *et al.* Universal features of post-transcriptional gene regulation are critical for *Plasmodium* zygote development. *PLoS Pathog.* **6,** e1000767 (2010).
29. Ponzi, M. *et al.* Egress of *Plasmodium berghei* gametes from their host erythrocyte is mediated by the MDV-1/PEG3 protein. *Cell. Microbiol.* **11,** 1272 (2009).
30. Orr, R. Y., Philip, N. & Waters, A. P. Improved negative selection protocol for *Plasmodium berghei* in the rodent malarial model. *Malar. J.* **11,** 103 (2012).
31. Janse, C. J., Ramesar, J. & Waters, A. P. High-efficiency transfection and drug selection of genetically transformed blood stages of the rodent malaria parasite *Plasmodium berghei. Nature Protocols* **1,** 346–356 (2006).
32. Sinden, R. *Molecular Biology of Insect Diseases Vectors: a Methods Manual* (eds Crampton J. M., Beard, C. B. & Louis, C.) (Chapman and Hall, 1997).
33. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina M-Seq sequensors. *BMC Genomics* **13,** 341 (2012).
34. Zerbino, D. R. & Birney, E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
35. Swain, M. T. *et al.* A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes. *Nature Protocols* **7,** 1260–1284 (2012).
36. Assefa, S. *et al.* ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25,** 1968–1969 (2009).
37. Tsai, I. J., Otto, T. D. & Berriman, M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.* **11,** R41 (2010).
38. Otto, T. D. *et al.* Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26,** 1704–1707 (2010).
39. Otto, T. D., Dillon, G. P., Degrave, W. S. & Berriman, M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res.* **39,** e57 (2011).
40. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
41. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
42. Campbell, T. L., De Silva, E. K., Olszewski, K. L., Elemento, O. & Llinás, M. Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathog.* **6,** e1001165 (2010).
43. Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnol.* **24,** 1429–1435 (2006).
44. Berger, M. F. & Bulyk, M. L. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature Protocols* **4,** 393–411 (2009).
45. Workman, C. T. *et al.* enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res.* **33,** W389–392 (2005).
46. Pfander, C., Anar, B., Brochet, M., Rayner, J. C. & Billker, O. Recombination-mediated genetic engineering of *Plasmodium berghei* DNA. *Methods Mol. Biol.* **923,** 127–138 (2013).
47. Zhang, Y., Buchholz, F., Muyrers, J. P. & Stewart, A. F. A new logic for DNA engineering using recombination in *Escherichia coli*. *Nature Genet.* **20,** 123–128 (1998).
48. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/ (2012).
49. Smyth, G. K. in: *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (eds Gentleman, R., Carey, V., Dudoit, S., Irizarry, R. & Huber, W.) 397–420 (Springer, 2005).