doi:10.1038/nature12970

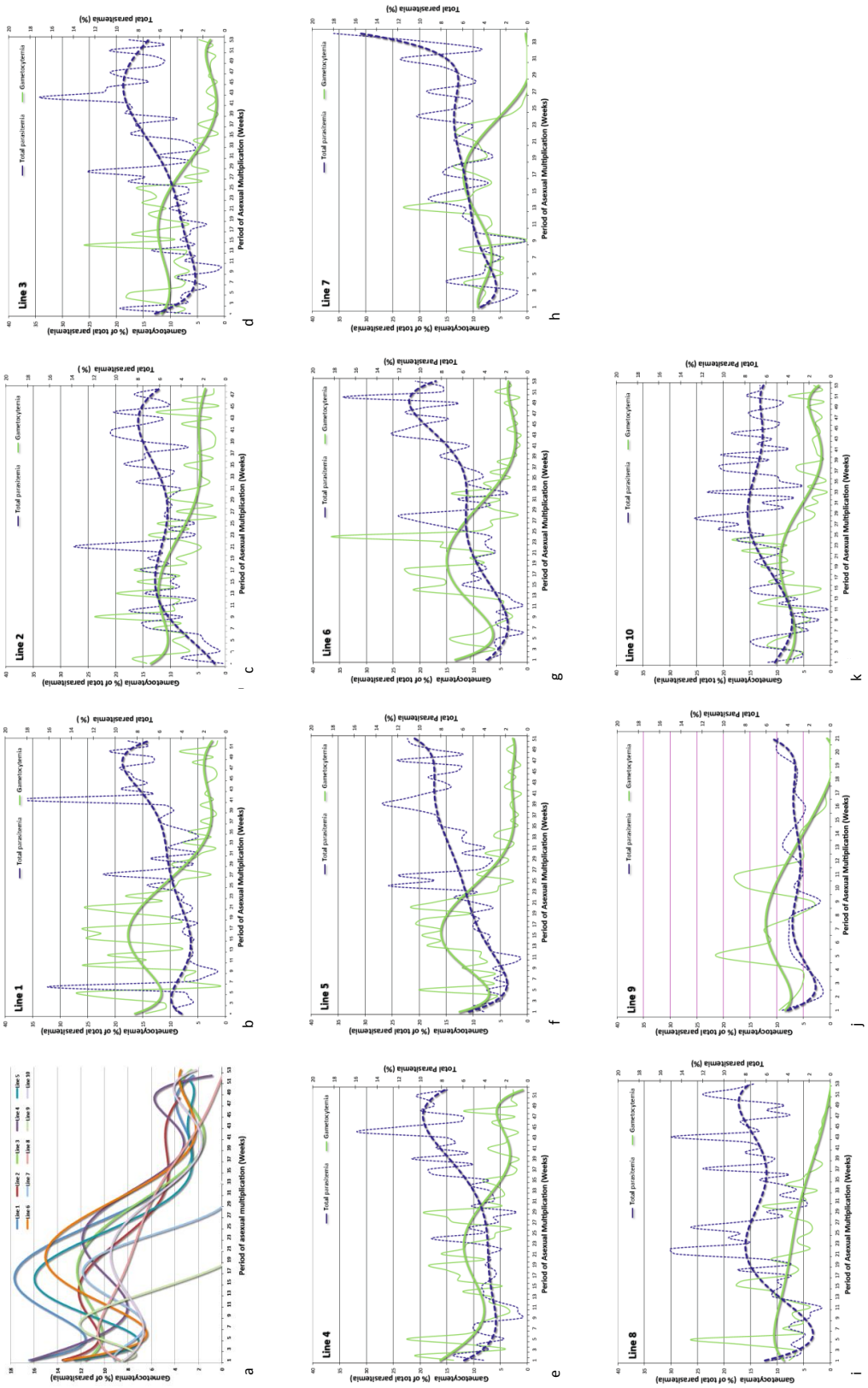


Figure S1. Parasitological parameters of continuous mechanical passage of 10 lines of PBANKA line 820 m1 cl1.

a. Gametocytaemias of the 10 lines expressed as a percentage of parasitaemia measured on a weekly basis by FACS as described in Materials and Methods. **b-k.** Parasitological parameters of the individual lines indicating best fit polynomial lines (bold) of the weekly readings (thin lines) for the gametocytaemia (green) and the parasitaemia (blue dashed).

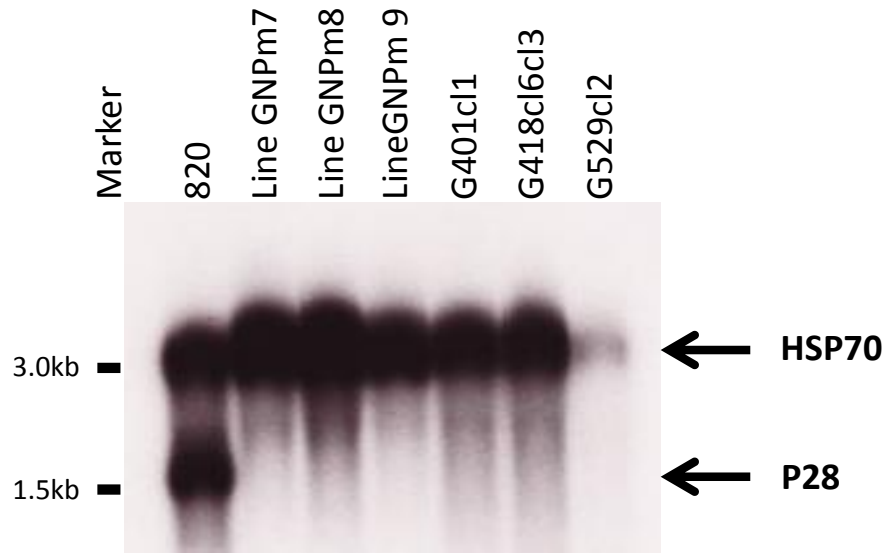


Figure S2: GNP and *ap2-g* KO lines do not produce gametocytes
Northern blot of total RNA from lines 820 (WT line), GNP lines with naturally acquired mutations (Lines m7-9) and *ap2-g* knockout mutant lines G401cl1 (full orf knockout, referred to as *ap2-g* KO1 in the manuscript), G418cl6cl3 (DNA binding domain knockout) and G529cl2 (partial ORF knockout eliminating the region with the 3 naturally acquired GNP mutations). The blot was probed for sexual stage specific transcripts with a ^{32}P labelled *p28* dsDNA (~1.5 kb band in WT line 820 and no bands in GNP lines) and normalized with a *hsp70* probe (~3 kb band in all lines). The absence of *p28* signal from the all GNP lines is consistent with the absence of sexual stages in the GNP lines. Samples tested individually and publication blot then performed.

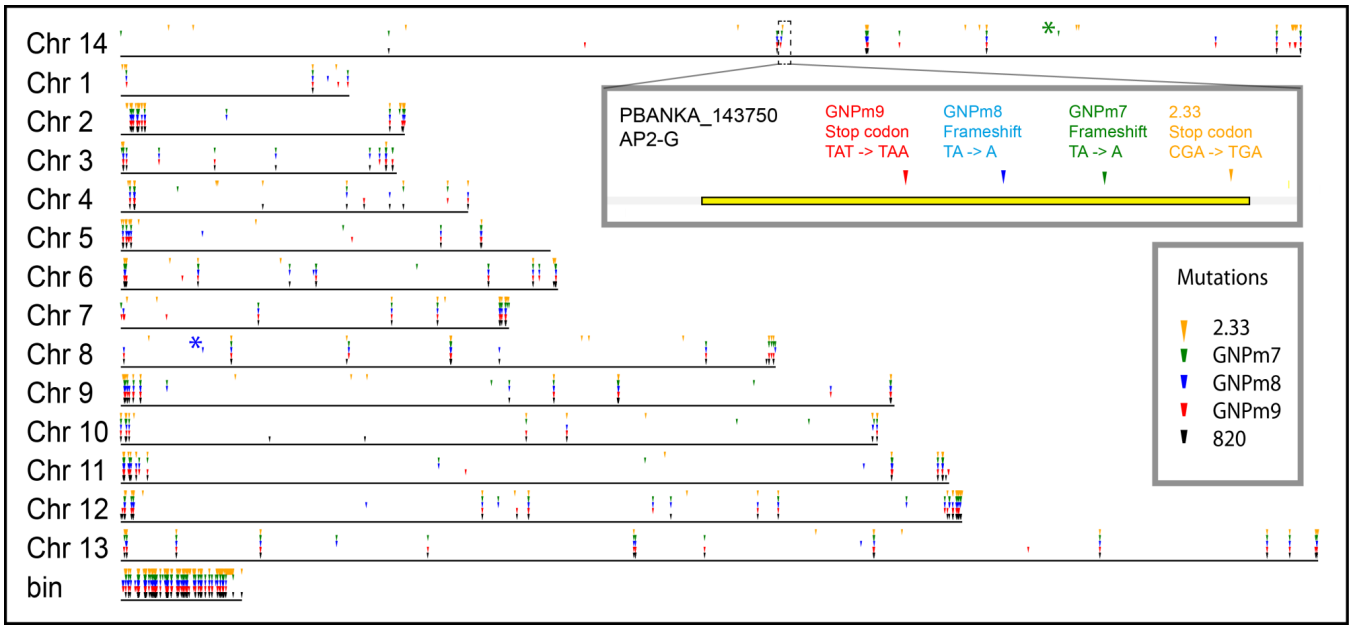


Figure S3. Distribution of potential mutations across all lines sequenced in this study. The location of all variant calls (see Supplementary Materials and Methods) are shown based on mapping to the assembled *P. berghei* 820 genome. Mapping SNPs in reads from clone 820 against their assembly reveals that the majority of SNPs are incorrectly called most likely due to assembly errors in low complexity genes. Many of these false calls are therefore shared by line 820 and GNP clones. From the initial variant calls, only one gene (*ap2-g*; see box on chromosome 14) contained variant sites in all three derived lines (GNPm7-m9) but not the parent (820). An overview of all variant calls and the filtering process that led to *ap2-g* is given in Table S3. The bin indicates unordered contigs not associated with chromosomes and is highly enriched in low complexity sequence and large multigene families (see Supplementary Materials and Methods). Asterisks indicate the chromosomal location of the only two additional genes that contained "homozygous" mutations (PBANKA_080320 in GNPm8 and PBANKA_145190 in GNPm7).

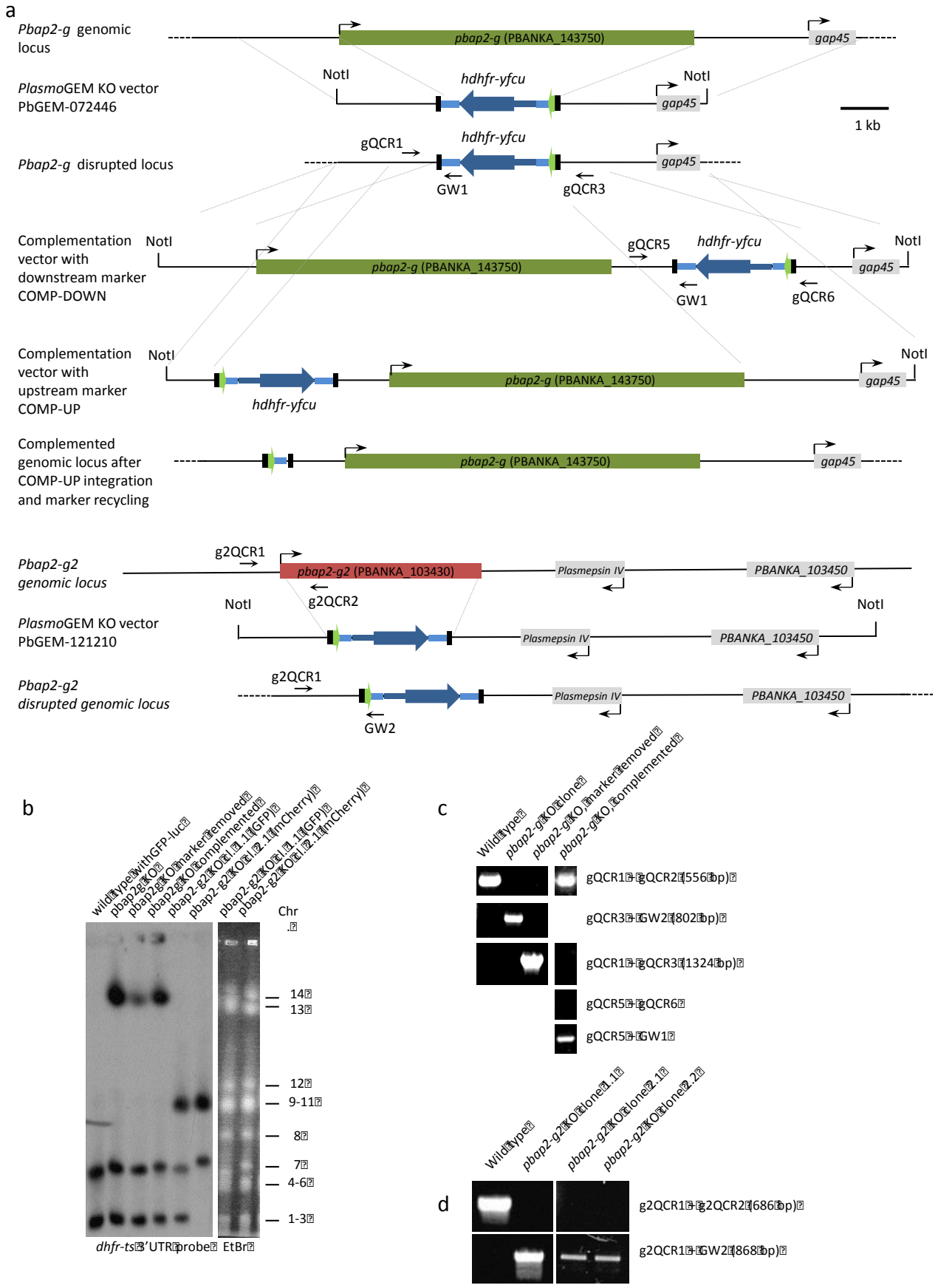


Figure S4a-d. Production and deployment of *Plasmo*GEM vectors

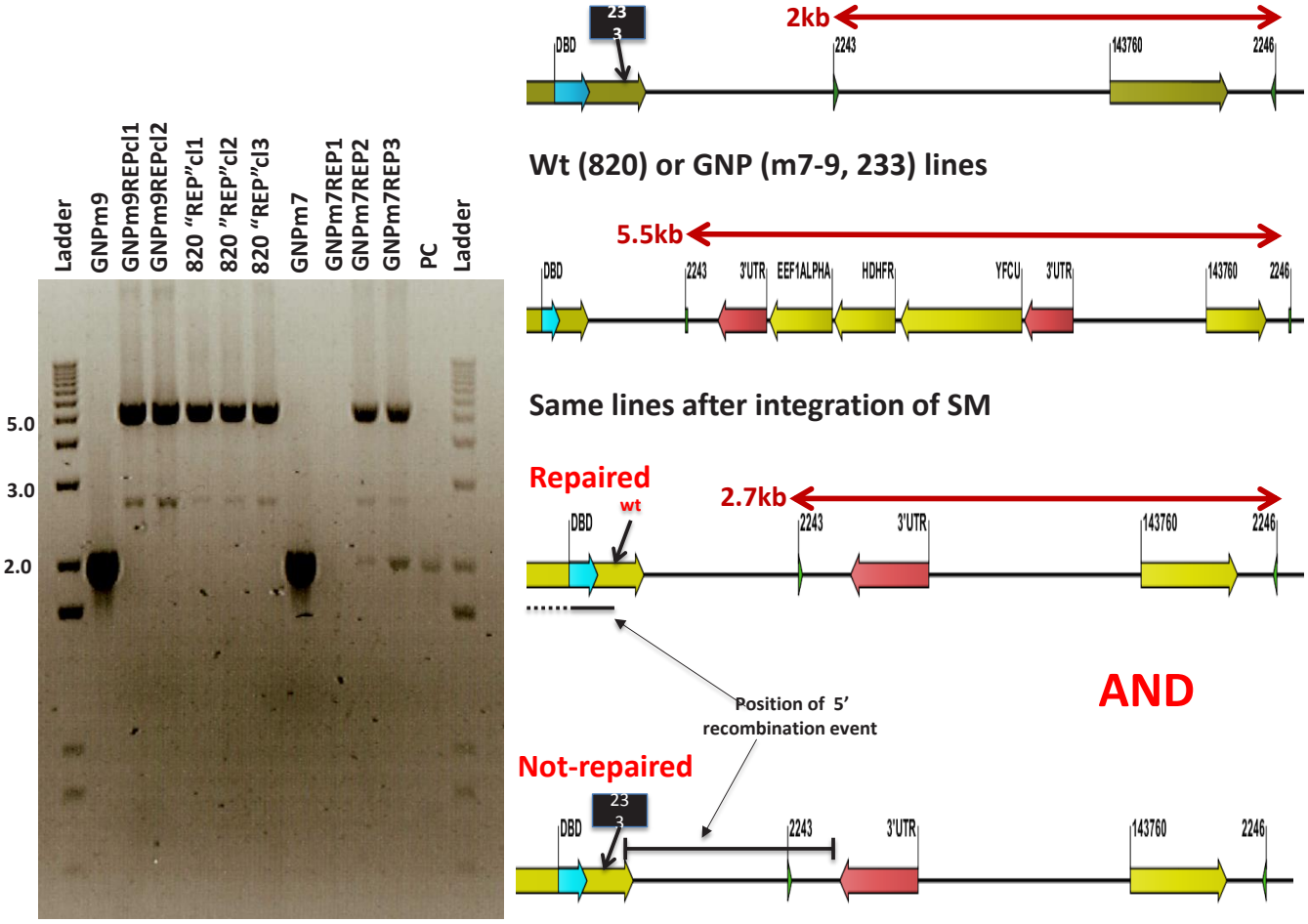
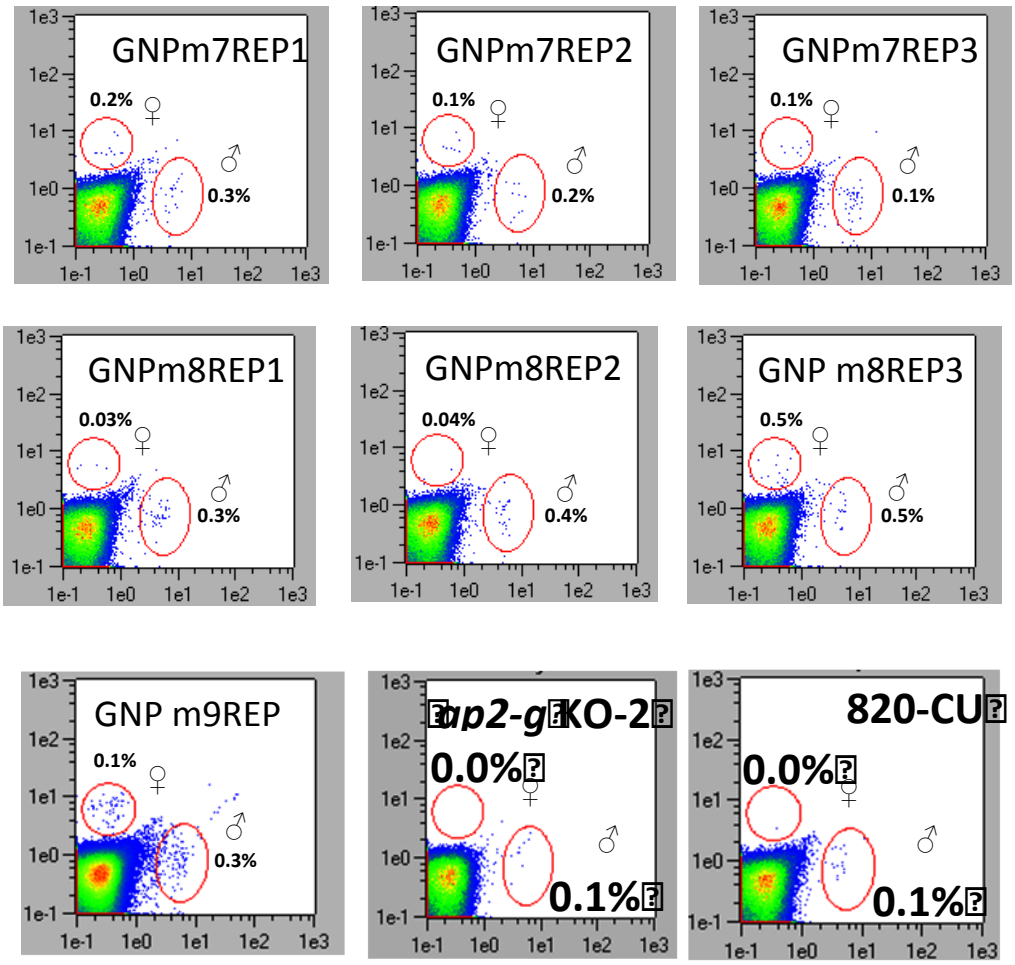


Figure S4e. Southern analysis and cartoons of the unmodified and repaired lines

Figure S4f. Quantitation of gametocytogenesis in lines produced in this study



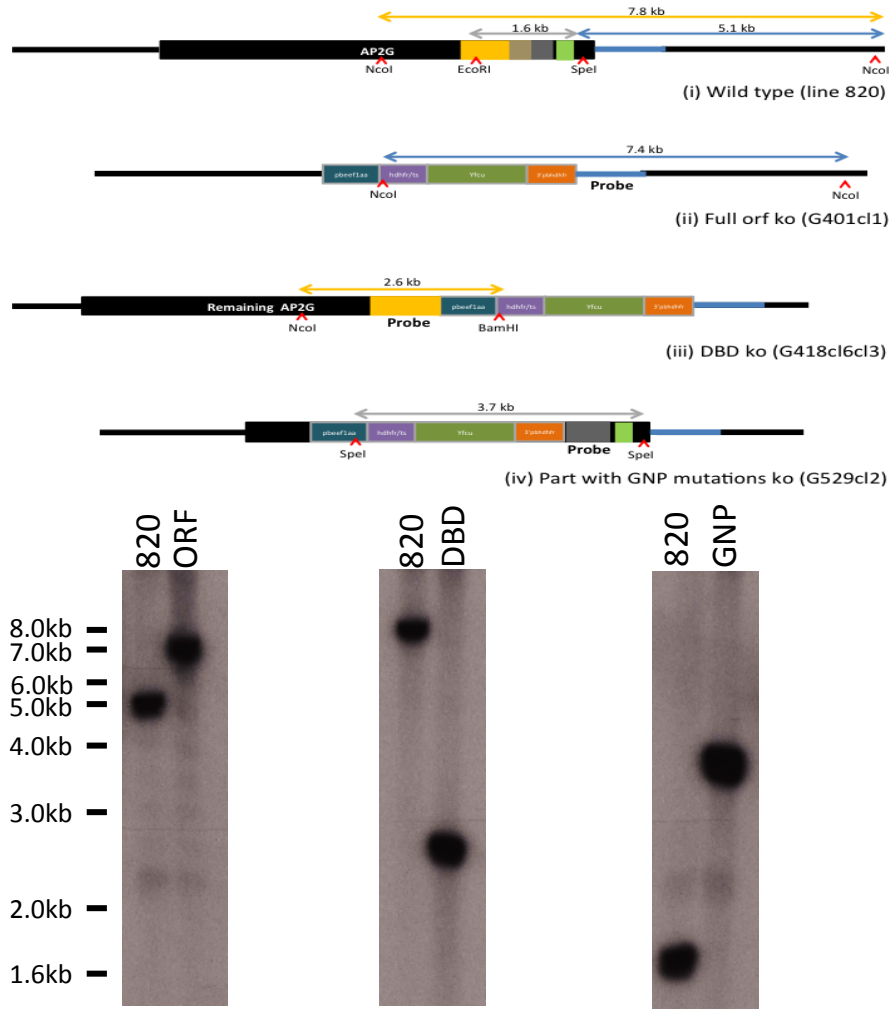


Figure S4g. Vectors used to disrupt *ap2-g* and Southern analysis of their integration into the *P. berghei* genome

Figure S4 Generation of knock out and complementation vectors used in this study and characterization of the subsequent parasite lines/clones generated following genetic transformation

a. Schematic of *Plasmo*GEM vectors for the targeted deletion and complementation of *pbap2-g* and *pbap2-g2* made by recombinase mediated engineering of *Plasmo*GEM genomic DNA library clones. Following targeted deletion of *pbap2-g* (resulting in *pbap2-g* KO2), the *dhfr-yfcu* selection cassette was removed by negative selection against *yfcu*. Complementation vectors carrying a selection marker either in the upstream or the downstream intergenic region (exact location as indicated in base pairs from the *pbap2-g* start codon) were then introduced into marker-free KO clones. Only the construct with the selection cassette downstream of *pbap2-g* restored gametocyte formation. **b.** Southern hybridization of separated chromosomes from the different *pbap2-g* and *pbap2-g2* mutants illustrated in panel S4A (n=1). A 500 bp fragment from the 3'UTR of the *P. berghei dhfr-ts* gene was used as a probe. Integration of the *pbap2-g2* KO vector on chromosome (chr.) 14 (lane 2) and of the *pbap2-g* KO vector in chr. 10 (lanes 5 and 6) introduces two copies of the probe sequence into these chromosomes. Marker recycling in the *pbap2-g2* mutant removes one target sequence for the probe on chr. 14, reducing band intensity (lane 3). The

second copy is restored after complementation (lane 4). The probe additionally reveals the endogenous *dhfr-ts* locus on chr. 7 and the GFP-luciferase expression cassette on chr. 3 (lanes 1-5). The last lane of the blot shows an independent *pbap2-g2* clone generated in a different genetic background expressing mCherry. Ethidium bromide staining of resolved chromosomes for two of the samples are shown (PCRs performed three times minimum). **c.** Diagnostic PCR verifying the genotype of the *pbap2-g* KO clone in panel S4a, removal of the selection marker and insertion of the complementation cassette with the marker in the downstream intergenic region. Oligonucleotides used are illustrated in panel S4a (PCRs performed three times minimum). **d.** Genotypes of three *pbap2-g2* KO clones generated on two different genetic backgrounds verified by PCR with primers shown in panel S4a (PCRs performed three times minimum). **e.** Southern analysis and cartoons of the unmodified and repaired lines (also following negative selection to remove the selectable marker) of the GNP lines derived from *P. berghei* ANKA line 820 m1 cl1. GNPm9REP clones (1 and 2, experiments G655-6); 820REP clones (1-3, experiments G657-9); GNPm7REP lines (experiments G696-8); PC = plasmid construct. NB GNPm7REP1 (G696) had insufficient DNA and no signal is observed. The cartoons also indicate the potential outcomes of an attempted repair which are dependent on the site of 5' recombination of the recombineering vector relative to the point mutation to be corrected. If recombination occurs upstream of the site for modification the correction will occur. If it occurs downstream the original mutation will remain unrepaired. SM = Selectable Marker. **f.** Quantitation of gametocytogenesis in lines produced in this study. Restoration of gametocytogenesis in GNPm7, m8 and m9 using pJazz based vector COMP-DOWN creating GNPm7REP (top row), GNPm8REP (middle row) and GNPm9REP (bottom row, left) respectively. For more detail see legend to Figure 1, Table S5 and Materials and Methods. FACS analyses of three independent transfections are illustrated for GNPm7REP and GNPm8REP. Disruption of gametocytogenesis in line 820 through whole gene deletion using the recombineering vectors described in Figure S4A, PbGEM-072446 (*ap2-g* KO2, bottom row, centre) and COMP-UP (820-CU, bottom row, right) is also shown. **g.** Southern blot: Schematic (not to scale) showing the restriction digestion pattern of gDNA from AP2-G full orf knockout (ii), AP2-G DBD knockout (iii) and AP2-G part ORF bearing the GNP mutations 7, 8,9 knockout (iv) double-digested with NcoI/SpeI, NcoI/BamHI and EcoRI/SpeI, respectively. The restriction pattern of WT 820 line being digested with all the 4 enzymes is shown in (i). The red arrowheads denote the site of action of the respective RE's. Probes for the three blots are indicated and were used to probe Southern blots of: NcoI/SpeI digested gDNA (left) from 820 WT (5.1kb) and G401cl1 (7.4kb); NcoI/BamHI digested gDNA (centre)from 820 WT (7.8kb) and G418cl6cl3 (2.6kb); and EcoRI/SpeI digested gDNA (right) from 820 WT (1.6kb) and G529cl2 (3.7kb) (samples tested individually and then publication blot performed).

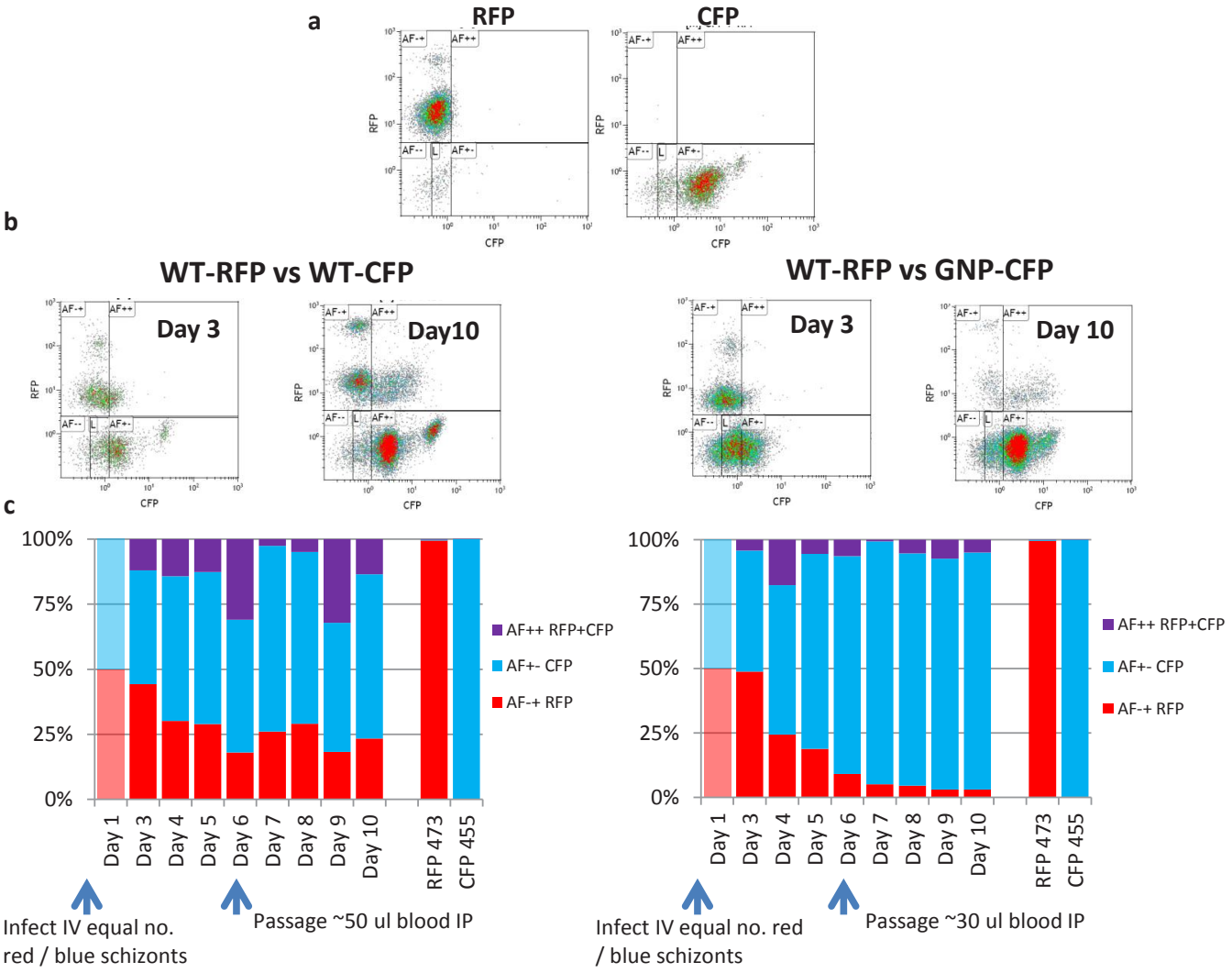


Figure S5. Competitive outgrowth of a non-producer line versus wild type parasites

a. GNPm9M1Cl1 was transfected with construct pG0148 (see reporter supplementary information for description) 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). b. Representative FACS plots after gating on infected cells showing CFP vs. RFP. Gates are indicated. c. Parasites were monitored daily by flow cytometry and after gating for infected cells the percentage of the population expressing either RFP (red), CFP (blue) or both (purple), reflecting mixed-multiply infected cells, was calculated and plotted. Calculated percentage in each gate is 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.

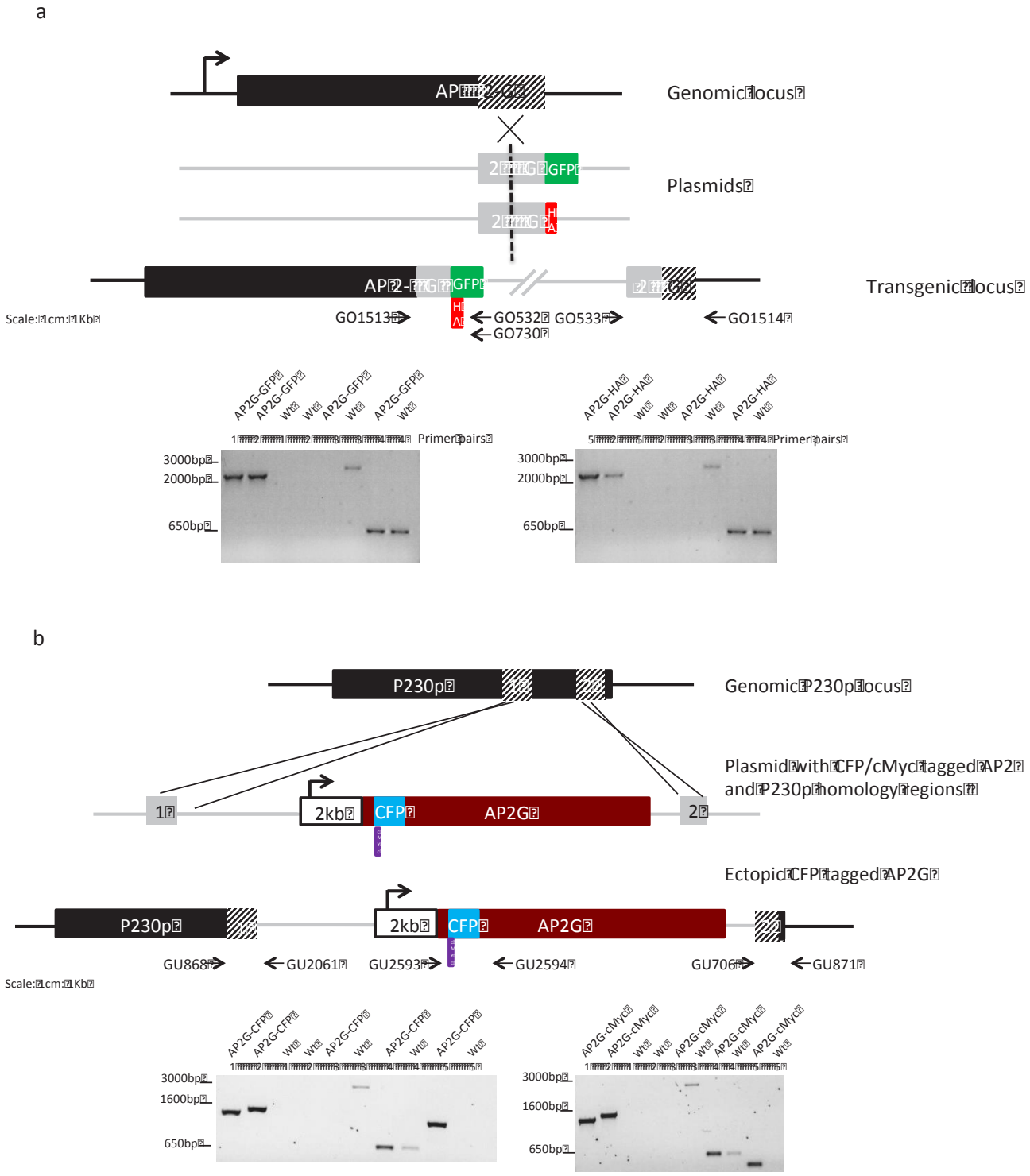


Figure S6a, b. Epitope tagging of *pbap2-g*

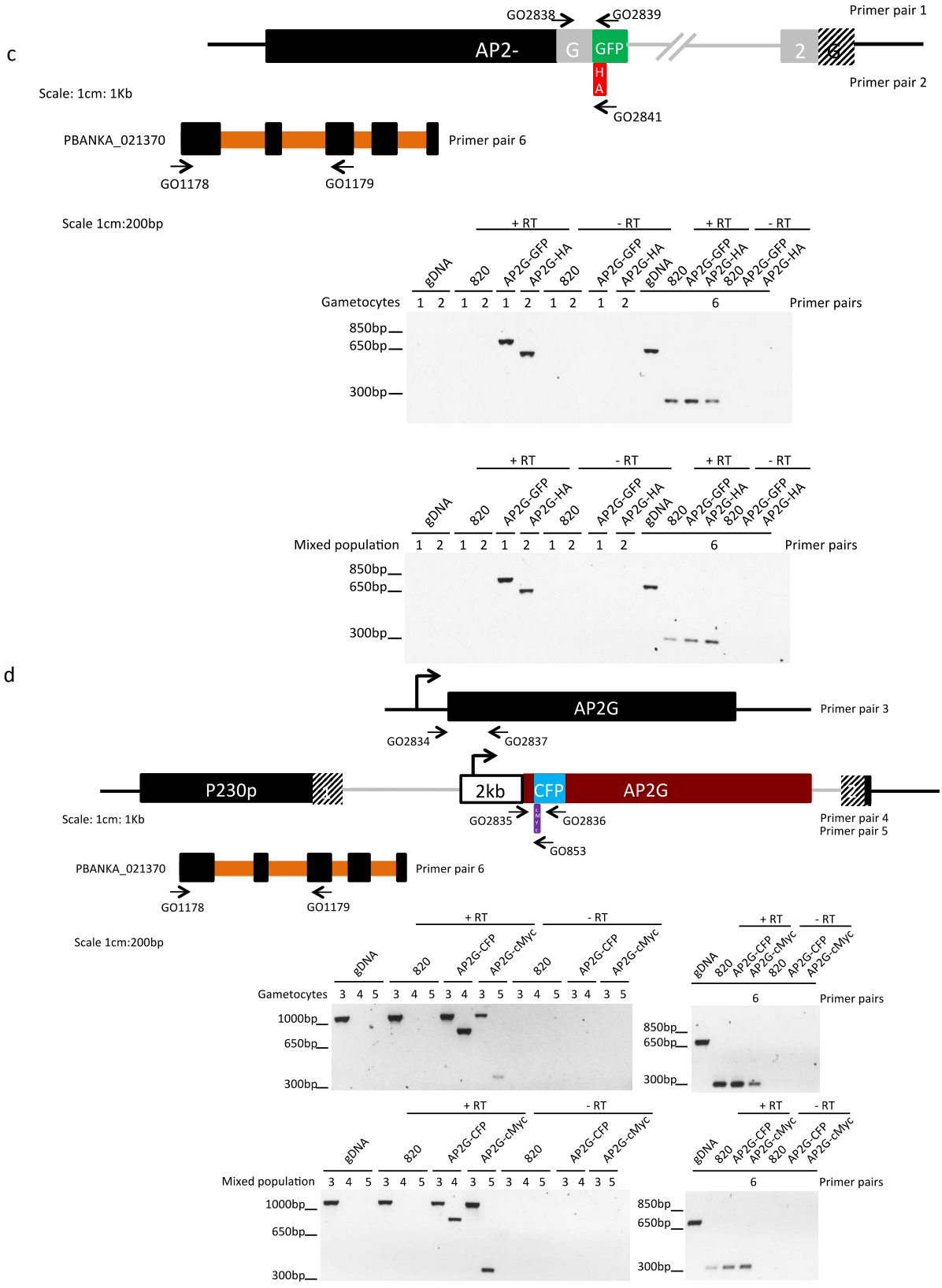


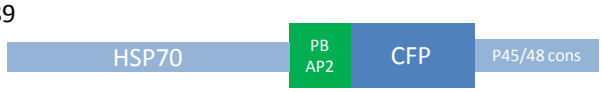
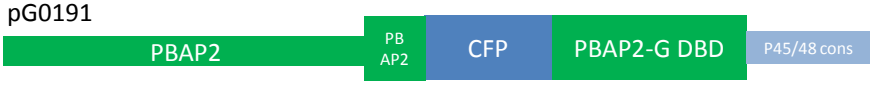
Figure S6c, d. Tagged *pbap2-g* is expressed in transfected *P. berghei*.

Figure S6 Tagged *pbap2-g* is expressed in transfected *P. berghei*.

a. Schematic representation of C-terminal tagging vectors for *ap2-g* showing the endogenous and the modified gene locus. Tagging of *ap2-g* utilized either a GFP (line G441)

or HA (line G443) tag via a single cross over event, (SXO.) The gene was linearized by the unique restriction enzyme XbaI, which was engineered into the gene via overlap extension. Diagnostic PCR confirmed successful integration using the following primers: 1. GO1513/GO532 (2.2Kb), 2. GO533/GO1514 (2.2Kb), 3. GO1513/GO1514 (2.6Kb), 4.GO13/GO14 (640bp covering the ORF of Pb28 PBANKA_051490.), 5. GO1513/GO730 (2.2Kb.). **b.** Schematic representation of *ap2-g* internally tagged vectors showing endogenous and modified 230 gene locus. Tagging of *ap2-g* utilized either a CFP (line G821) or cMyc (line G872) tag via a double cross over event, (DXO). The *ap2-g* gene was optimized for *E. coli* codon bias and therefore has a different nucleotide sequence from the endogenous *ap2-g* gene. Diagnostic PCR confirmed successful integration using the following primers: 1. GO868/GO2061 (1.43Kb), 2. GO706/GO871 (1.53Kb), 3. GO868/GO871 (2.6Kb), 4. GO13/GO14 (640bp covering the ORF of *p28,* PBANKA_051490.), 5. GO2593/GO2594 (1.2Kb CFP/ 470bp cMyc.). **c**. Schematic representation of *ap2-g* C terminally tagged modified genes and PBANKA_021370 control gene with introns shown in orange. RNA was extracted from these GFP- (G441) and HA- (G443) modified *ap2-g* parasite lines and used to generate cDNA. As *ap2-g* has no introns, PBANKA_021370 was used as a second control to RT –ve treated RNA. Diagnostic PCR confirmed successful transcription using the following primers: 1. GO2838/GO2839 (875bp), 2. GO2838/GO2841 (650bp), 6. GO1178/GO1179 (733bp gDNA/310bp cDNA). 820 wt cDNA was further used as a control to show integration of the GFP/HA tags. **d.** Schematic representation of *ap2-g* internally tagged modified genes and PBANKA_021370 control gene with 4 introns shown in orange. RNA was extracted from these CFP- (G821) and 2xcMyc- (G872) modified AP2G parasite lines and used to generate cDNA. As *ap2-g* has no introns, PBANKA_021370 was used as a second control to RT –ve treated RNA. As these internally tagged modified genes are at an ectopic locus, primers have been used to detect both endogenous and ectopic transcripts. Diagnostic PCR confirmed successful transcription using the following primers: 3. GO2834/GO2837 (1.15Kp), 4. GO2835/GO2836 (860bp), 5. GO2835/GO853 (365bp) 6. GO1178/GO1179 (733bp gDNA/310bp cDNA). 820 wt cDNA was further used as a control to show integration of the GFP/HA tags and a single endogenous transcript (all PCRs performed three times minimum).

a

CONSTRUCT	820	GNP9
pG0148 	Blue All stages	Blue All stages
pG0157 	No detectable blue	No detectable blue
pG0189 	Blue All stages	Blue All stages
pG0190 	Lethal?	Lethal?
pG0191 	Nuclear localised Blue in Females (and ooks?)	No detectable blue

b

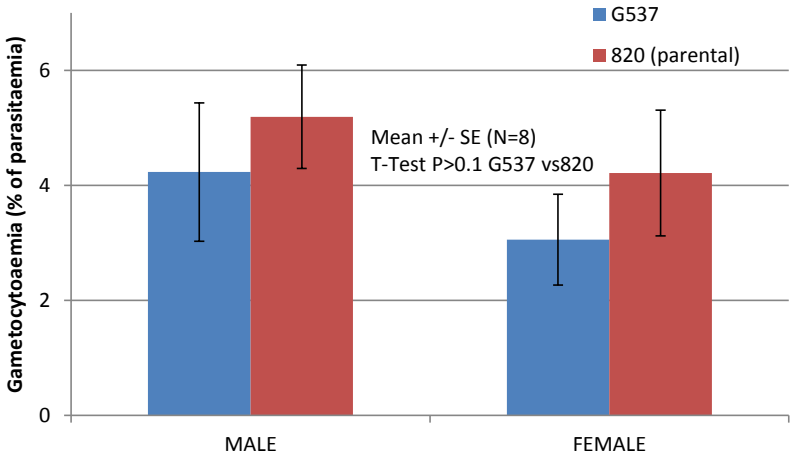


Figure S7. Expression characteristics and control experiments of the *ap2-g* minigene
a. Summary of the analysis of the expression of ap2-g minigene constructs. The C terminal 900 bp of the pbap2g appears to be sufficient for nuclear localisation, whereas the N terminal 300 bp conferred no apparent localisation to CFP. No parasites were observed in lines transfected with pG0190 suggesting the huge overexpression of the nuclear targeted construct achieved by HSP70 promoter is lethal. The level of CFP expression where observed in pG0191 transfected constructs was very low and not detectable by flow cytometry, The enrichment or stabilisation of signal by localisation to the nucleus was the only way expression driven by the pbap2g promoter could be detected suggesting it drives low levels of expression. No change in the number of gametocytes could be seen after transfection of 820 with pG0191 and no rescue of GNPm9 was seen suggesting this minigene is not functional. **b.** Bar chart showing mean gametocytaemia from 8 experiments +/- SE showing that expression of the minigene in the nucleus of line 820 has no effect on gametocyte production. P values from 2-tailed t-test for Males G537vs820 = 0.559 and Females G537vs820=0.436.

a

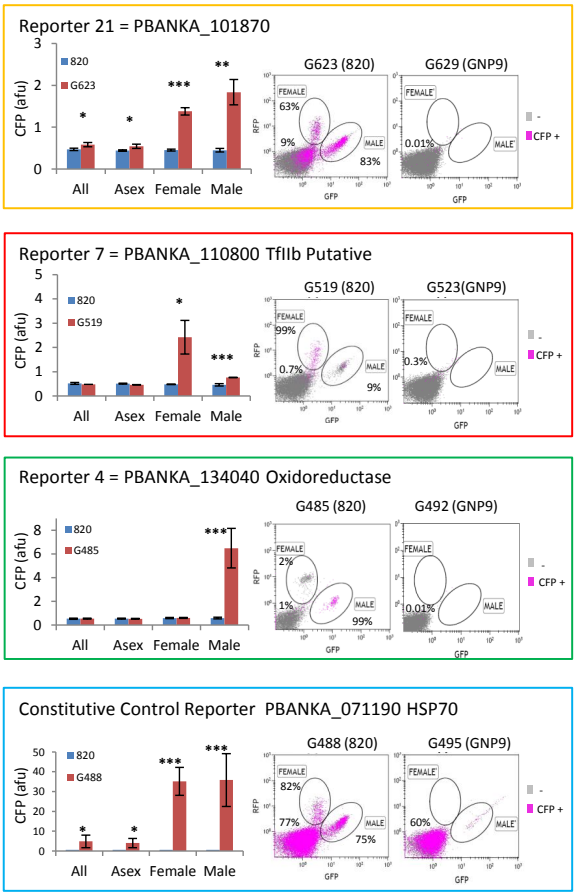


Fig S8a Life cycle specific CFP expression from selected reporter promoters

b

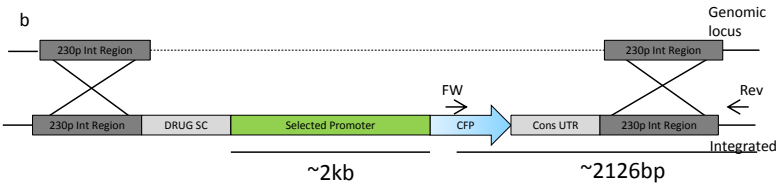


Fig S8b Schematic of integrated reporter constructs

c

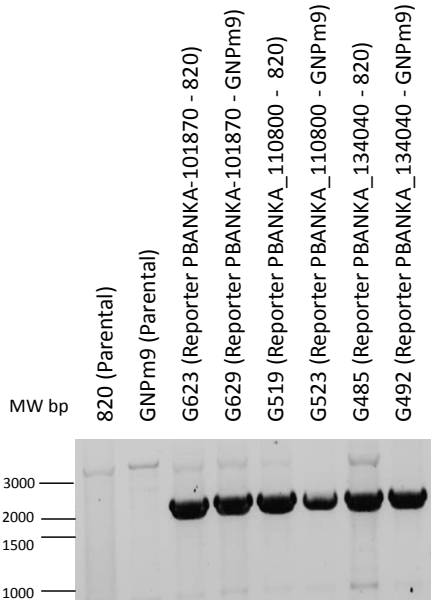


Fig S8c PCR results showing integration of reporter constructs

d

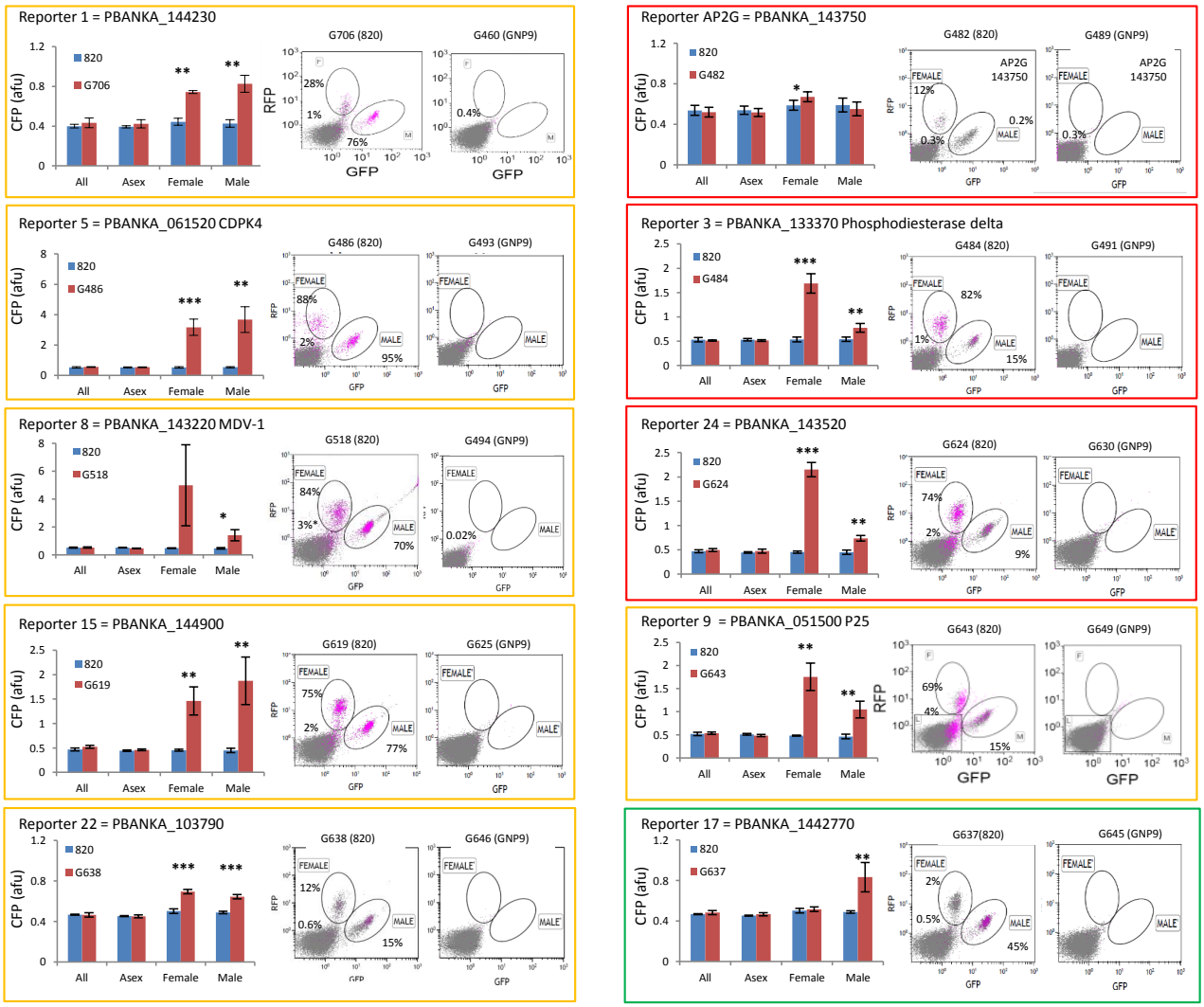


Figure S8. Identification and characterisation of promoters of gametocyte specific genes.
a. Bar plots show geometric mean CFP expression for developmental stages separated by FACS gating according to GFP or RFP fluorescence in parental (820) line (blue bars) and the reporter line made in the GNPm9 parental background (red bars). Mean of three independent measurements +/- SD. P value from 2-tailed T-test: *<0.05 **<0.01 ***<0.001. Flow cytometry plots show a scatter plot of all infected cells (gated on DyeCycle Ruby DNA stain) plotted GFP (X-axis) vs. RFP (Y-axis) to illustrate male, female and asexual populations. Grey points show infected red blood cells, which are negative for CFP and magenta points show infected red blood cells which are additionally CFP positive. Left plot shows reporter in 820 (parental) background and right plot shows reporter in GNPm9 background. Numbers show the percentage of the population within each gate which are CFP positive. Flow cytometry plots show one example representative of three experiments. Expression of CFP (magenta) specifically in gametocyte stages in 820 background illustrates stage specificity driven by the promoter sequence, and absence of CFP expression in GNPm9 confirms lack of gametocyte specific transcripts in GNP lines as predicted by microarray data. In total 18 reporters were analysed (Table S8). The few events falling into the male GFP positive gate in the hsp70 reporter in GNPm9 background are probably a result of bleed through from the

bright CFP expression driven by hsp70 promoter in some parasites as this was not seen with any other line. Reporter PBANKA_101870 was identified through down regulation in trophozoite stages only (GNP vs. 820) and appears to be an example of an early expressed male and female gametocyte gene. PBANKA_134040 is male specific, PBANKA_110800 shows specifically higher female expression. The constitutive control hsp70 shows expression in both the 820 line and the GNPm9 line showing that these are both capable of expressing reporter constructs. **b**. Schematic diagram (not to scale) of reporter construct integration by double crossover homologous recombination into the p230p locus and location of primers (FW and Rev) used to confirm construct integration. **c**. PCR results showing integration of CFP reporter constructs as illustrated in schematic S8b in both the parental (820) background and the GNPm9 background (n=1). **d**. Further examples of gametocyte specific reporters. Bar charts and Flow cytometry plots as described for FigS9a. Colour coding of borders indicates male (green), female (red), or both male and female (yellow) specific expression.

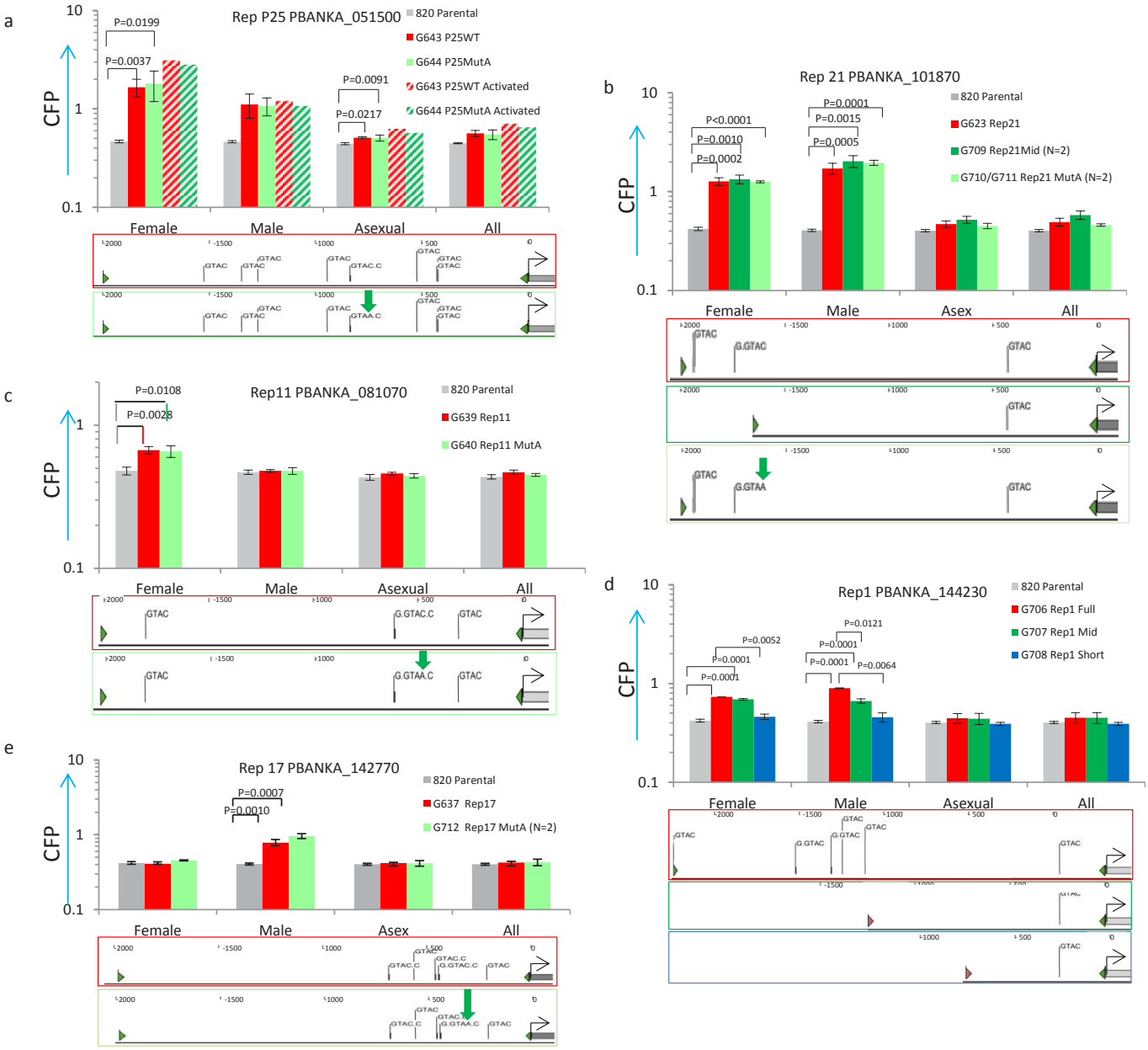


Figure S9. Characterisation of (mutated and truncated) reporter constructs in transfected *P. berghei*.

CFP expression driven by wild type and mutated reporter promoters. Each panel a-e illustrates a different promoter driving CFP expression expressed from a 230p targeted construct (see fig S8): PBANKA_051500 (a), PBANKA_101870 (b), PBANKA_144230 (c), PBANKA_081070 (d), PBANKA_142770 (e). Bar plots show CFP expression (geometric mean CFP expression level) driven by the 2kb of the respective promoter sequence. Grey bars show parental 820 (background) control CFP level. CFP expression from WT promoter is shown in (red) and CFP expression from 2kb of mutated expression (green). Mutations are point mutations in the AP2-G binding motif (G.GTAC or GTAC.C) to mutate the central T residue to an A. Mean +/- SD from three experiments, P values illustrate 2-tailed T test result. Life cycle stages asexual, male and female are separated based on GFP and RFP fluorescence as described earlier. Schematics below each bar plot indicate the promoter

region and location of the mutations is indicated by a green arrow. For b) and d) truncated promoters are also shown (dark green and blue bars) as illustrated in the schematics below. (N= 3 unless otherwise stated).

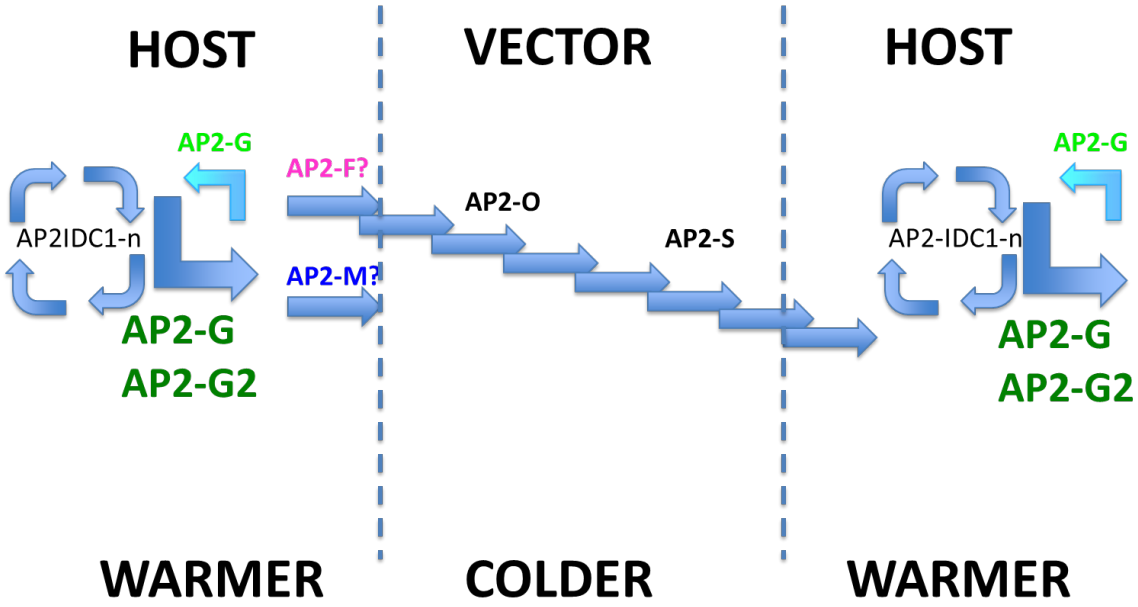


Figure S10. ApiAP2 transcription factor family and the *Plasmodium* life cycle.
Schematic demonstrating the roles of *pbap2-g* and *pbap2-g2* in the context of the *Plasmodium* life cycle indicating the known roles of other factors in life cycle progression and indicating the postulated factors that might control sexual differentiation subsequent to commitment to gametocytogenesis.

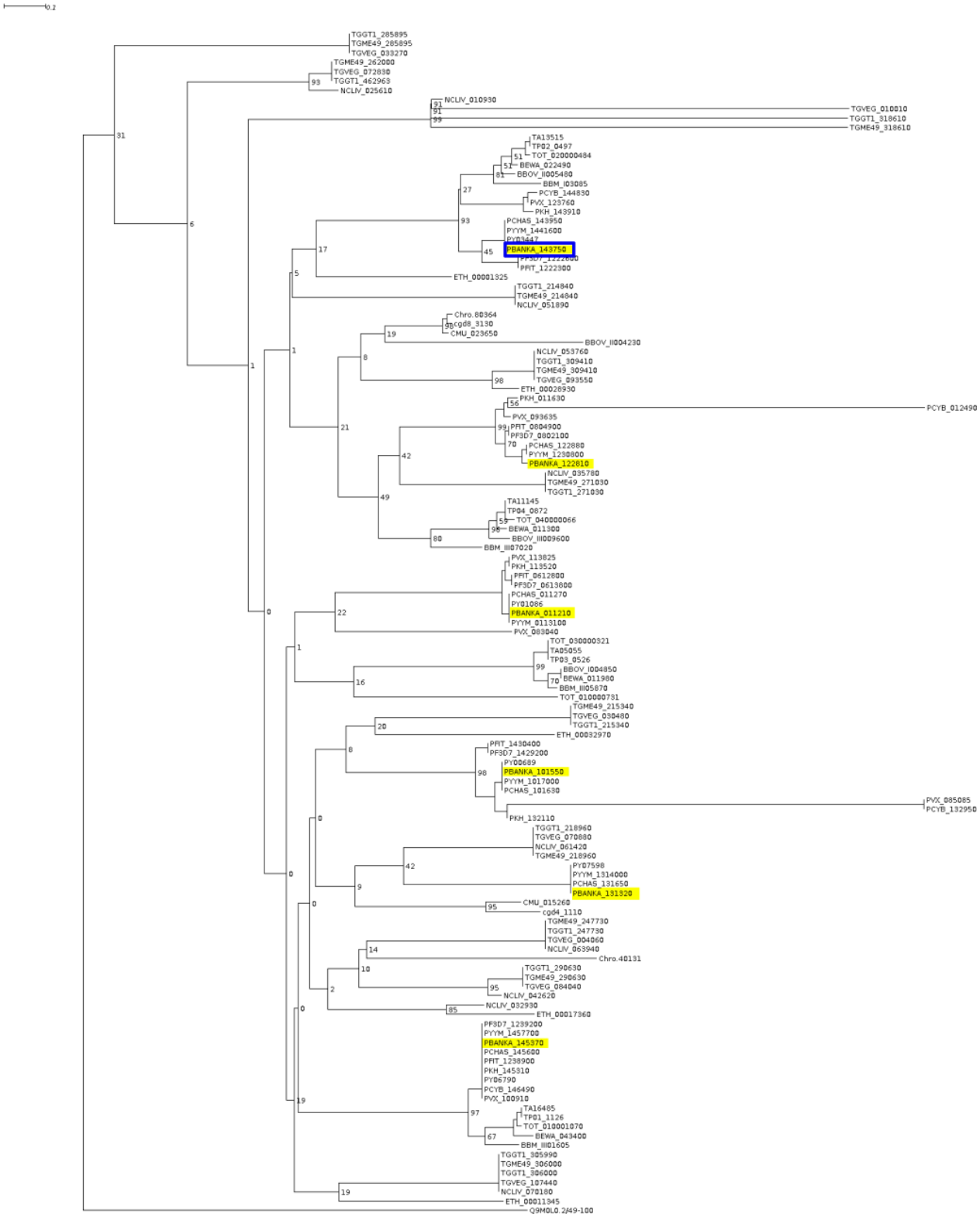


Figure S11. Distribution of PBAP2-G DBD in apicomplexan parasites.
Partial phylogeny of apicomplexan ap2-dbd reveals different DBDs form distinct clades and that coccidian DBDs cluster and group more distantly to those DBDs identified in *Plasmodium* and piroplasms. *P. berghei* DBDs are highlighted in yellow, PBAP2-G is bordered in blue.

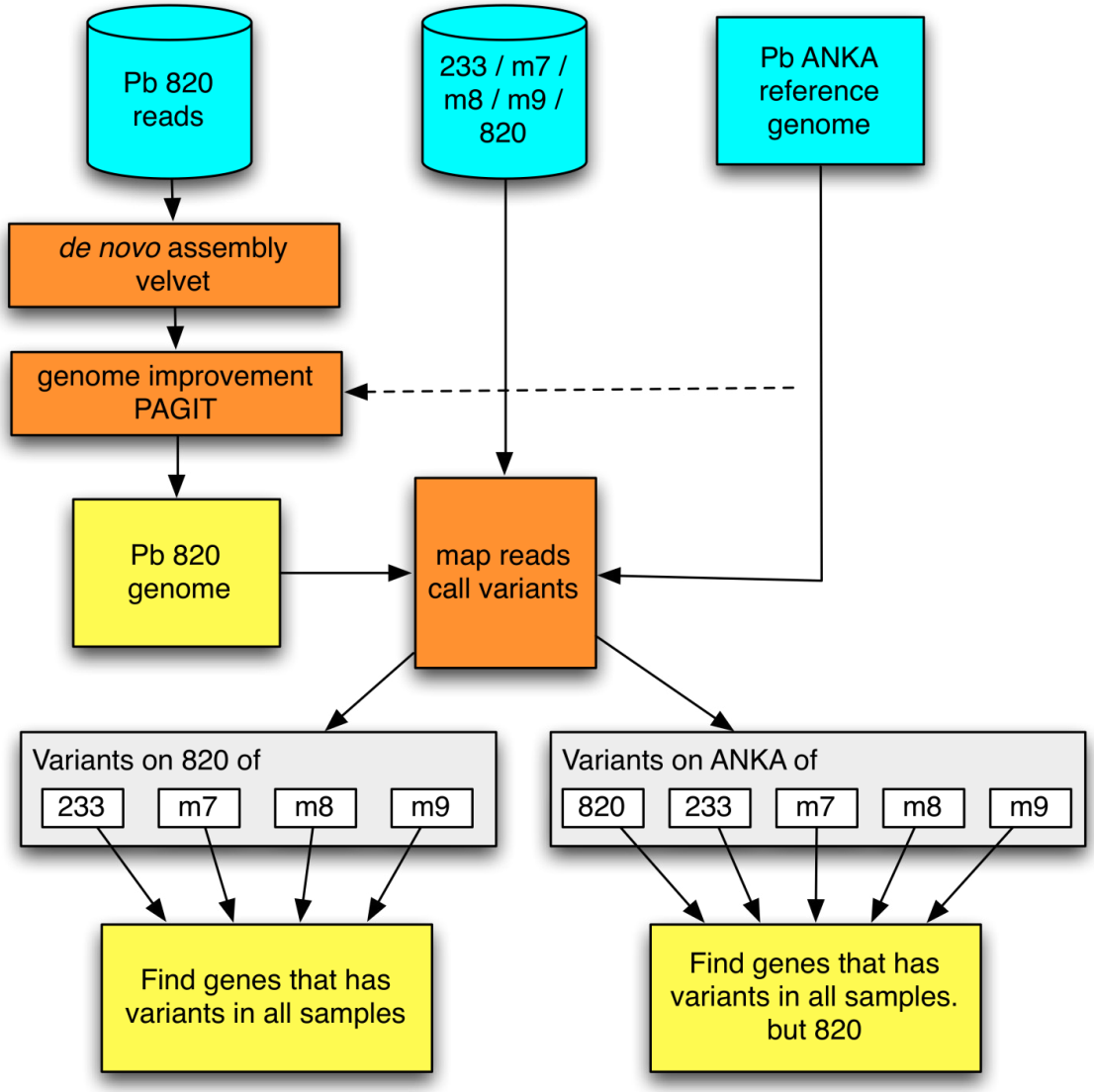
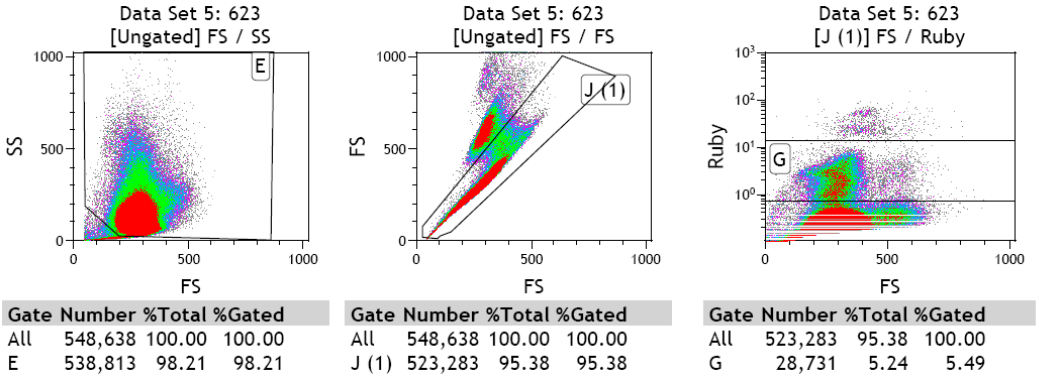


Figure S12. Pipeline for analysis of WGS of *P. berghei* GNP and 2.33 cloned lines

See Materials and Methods section for more detail

a

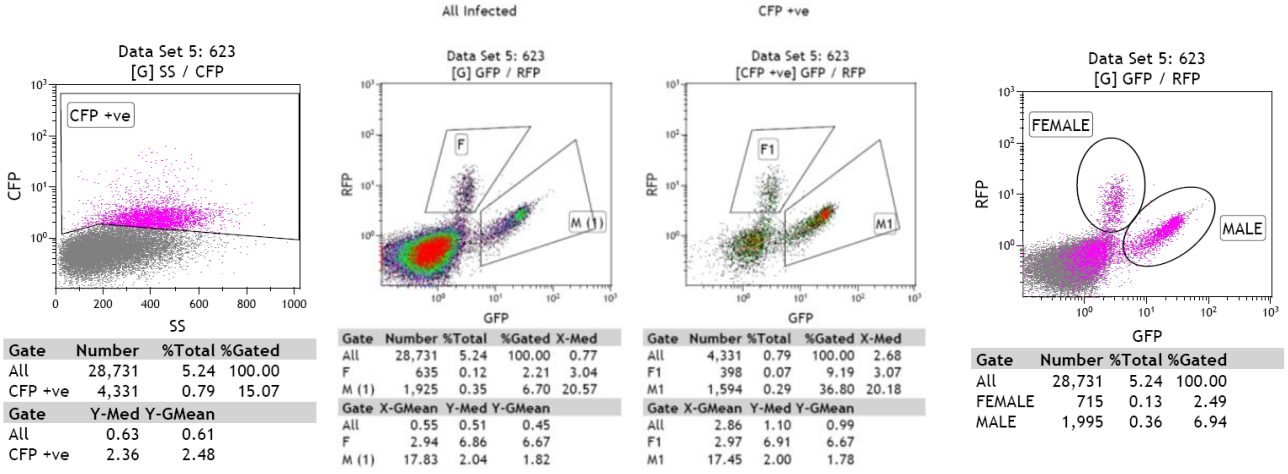


Gate	Number	%Total	%Gated
All	548,638	100.00	100.00
E	538,813	98.21	98.21

Gate	Number	%Total	%Gated
All	548,638	100.00	100.00
J (1)	523,283	95.38	95.38

Gate	Number	%Total	%Gated
All	523,283	95.38	100.00
G	28,731	5.24	5.49

b



Gate	Number	%Total	%Gated
All	28,731	5.24	100.00
CFP +ve	4,331	0.79	15.07

Gate	Y-Med	Y-GMean
All	0.63	0.61
CFP +ve	2.36	2.48

Gate	Number	%Total	%Gated	X-Med
All	28,731	5.24	100.00	0.77
F	635	0.12	2.21	3.04
M (1)	1,925	0.35	6.70	20.57

Gate	X-GMean	Y-Med	Y-GMean
All	0.55	0.51	0.45
F	2.94	6.86	6.67
M (1)	17.83	2.04	1.82

Gate	Number	%Total	%Gated	X-Med
All	4,331	0.79	100.00	2.68
F1	398	0.07	9.19	3.07
M1	1,594	0.29	36.80	20.18

Gate	X-GMean	Y-Med	Y-GMean
All	2.86	1.10	0.99
F1	2.97	6.91	6.67
M1	17.45	2.00	1.78

Gate	Number	%Total	%Gated
All	28,731	5.24	100.00
FEMALE	715	0.13	2.49
MALE	1,995	0.36	6.94

Figure S13. Gating strategy for FACS analysis of CFP expressing reporter lines

a. Gating strategy to identify parasite-infected erythrocytes stained with Vybrant DyeCycle Ruby. **b.** Gating strategy to demonstrate stage and gender specific expression of CFP reporter genes transfected into blood stage *P. berghei.* For detail see Materials and Methods.

PBANKA Gene ID	Gene model description	Reporter Number	Expression in 820	Expression in GNP9	Mutated Reporter
PBANKA_143750	AP2		F	None	
PBANKA_144230	Conserved Hypothetical	1	M+F	None	M+F
PBANKA_082670	Conserved Hypothetical	2	All	Low	
PBANKA_133370	Phosphodiesterase	3	F	None	
PBANKA_134040	Oxidoreductase	4	M	None	
PBANKA_061520	CDPK4	5	M+F	None	
PBANKA_110800	TFIIb	7	F	None	
PBANKA_143220	MDV-1	8	F+M	None	
PBANKA_051500	P25	9	F+M	None	F+M
PBANKA_081070	SPM-1	11	F	None	F
PBANKA_135970	6-Cys	13	F	None	
PBANKA_144900	Conserved Hypothetical	15	M+F	None	
PBANKA_050440	Conserved Hypothetical	16	F	None	
PBANKA_142770	DNA Helicase	17	M	None	M
PBANKA_051910	CCAT Transcription Factor	18	All	Yes	
PBANKA_140300	Small HSP	20	F+M	None	
PBANKA_101870	Conserved Hypothetical	21	M+F	None	M+F
PBANKA_103790	Conserved Hypothetical	22	F+M	None	
PBANKA_143520	Conserved Hypothetical	24	F	None	
PBANKA_071190	HSP70 (constitutive)		All	Yes	

Table S8. Summary of the reporter constructs employed in this study.
F = Female Gametocyte; M = Male gametocyte; All = all blood stage forms (asexual and gametocytes). Reporters which were used in the mutated or truncated promoter studies are indicated in bold for the gene ID

Protein	Old_id	DBD_ext MW (KDa)	Tag MW (KDa)	DBD::tag MW (KDa)
AP2-G DNA binding Domain	PFL_1085w (PF3D7_1222600)	14.7	26 (GST)	40.7
AP2-G DNA binding Domain	PBANKA_143750	14.3	26 (GST)	40.3

Table S10. Properties of the recombinant GST:DBD fusion proteins produced in this study.