Supplemental Figures and Tables legends Santos *et al*. 2017

Figure S1. Related to Figure 1.

A- Conservation of the "rhoptry" DNA sequence motif (in red) in invasion-related gene promoters from *P. falciparum* (Pf), *P. berghei* (Pb), *P. vivax* (Pv) or *P. knowlesi* (Pk). See also figure 1B.

B- Strategy used to C-terminally tag PfAP2-I with GFP. The presence of a PCR product with primers 1 and 4 or 2 and 4 only in the transgenic line demonstrates that the locus has been correctly modified. The primer sequences are listed in Table S1. See also figures 1D, E.

Figure S2. Related to Figure 1.

Protein sequence alignment of the AP2-I sequence of *P. falciparum* (Pf), *P. berghei* (Pb), *P. vivax* (Pv) and *P. knowlesi* (Pk). Boxed in dark grey is the ACDC domain, in dark blue are the AP2 domains, in light blue is the AT-hook domain and in light grey is the NLS. In red are highlighted the conserved amino acid residues. See also figure 1C.

Figure S3. Related to Figures 2 and 3.

A- ChIP-qPCR from a parasite line expressing PfAP2-I-GFP indicates that PfAP2-I does not associate with the *SPE2* motifs bound by PfSIP2. 3D7 *wt* and PfSIP2N-HA parasites were used as negative and positive controls, respectively. The results are shown as fold enrichment of ChIP performed with anti-GFP antibody versus ChIP performed with IgG (n=3). Data are represented as mean ± SD. See also figure 2D.

B- DNA motif analysis using DREME identified GTGCA as the most common DNA motif within all of the PfAP2-I ChIP-seq peaks (e-value=2E-25). The next five highest scoring motifs after GTGCA found in the ChIP-seq peaks are shown on the right. See Table S2 for a complete list of DREME results. Matching the top DREME motif against known DNA motifs bound by *Plasmodium* AP2 domains (Campbell *et al*., 2010) using TOMTOM shows that GTGCA matches the DNA motif bound by the third AP2 domain of PfAP2-I and the AP2 domain of PfSIP2 (lowest E-value) and has reduced similarity to other known AP2-associated sequence motifs. See also figure 3A.

C- Plot representing the distance of the top DREME DNA motif site relative to the ChIP-seq peak summits across the three ChIP biological replicates. The mean motif distance from the peak summit are +3, +2 and +3 bp for replicates 1-3.

D- Integrative genomics viewer (IGV) plot (Thorvaldsdottir *et al*., 2013) showing the genomic region encoding *msp2*, *msp4* and *msp5* and containing ChIP-seq trimmed peaks 3 and 4. The log2 enrichment ChIP/Input for each ChIP replicate is shown. A grey block represents the peak called by MACS2 (the number of the correspondent trimmed peak is indicated), arrows represent genes, and small black, red and blue lines represent the presence of a predicted DNA motif associated with PfAP2-I-D3 or –D2 or –D1, respectively.

E- Sanger sequencing verifying that the correct mutations were introduced in PfAP2-I-GFP::*msp5*MUT parasites, compared to the parental strain, PfAP2-I-GFP. Boxed in red are the mutated nucleotides. See also figure 3B.

F- ChIP-qPCR with PfAP2-I-GFP-*msp5*MUT parasites shows that PfAP2-I-GFP binding to *mps5* is diminished (p=0.07) (red box). PfAP2-I binding to other genes is not significantly altered (p>1) (AP2-I ChIP-seq targets). The results are shown as fold enrichment of ChIP performed with anti-GFP antibody versus ChIP performed with IgG (n=3). Data are represented as mean ± SD. 3D7 *wt* and PfAP2-I-GFP parasites were used as negative and positive controls, respectively. See also figure 3C.

G- Time points collected for the microarray timecourse experiment performed with PfAP2-GFP and PfAP2-I-GFP-*msp5*MUT parasites. See also Table S3 and Figure 3D.

H- Plot of the average Rnits (Sangurdekar, 2014) p-value for all transcripts detected by microarray (minus antigenic variant genes) showing that the *msp5* gene is one of the most significantly altered transcripts (average p-value 4.65E-2) (see Table S3 for full list of Rnits p-values). Rnits compares multiple time-series expression data sets by summarizing probes into gene-level information. For all genes, it fits a series of B-splines with varying curvature and degrees of freedom. Under the null hypothesis H_0, a single model is fit for all data sets. P-values from the hypothesis test are then plotted and the least complex spline parameters that result in uniformly distributed null p-values are automatically chosen. Each gene is attributed a p-value from 0 to 1 until all p-values are uniformly distributed. While genes with p-values closer to 1 have less changes in transcription levels between expression data sets, genes with p-values closer to 0 have higher transcription levels differences

between time-series expression data sets. The transcription level of *msp5* drops significantly in the mutant parasites, and is therefore attributed a low Rnits p-value.

I- RT-qPCR experiments show that while *msp5* transcript levels are reduced in PfAP2-I-GFP-*msp5*MUT parasites, the transcript levels of other genes are similar in *wt* and mutant parasites (PfAP2-I-GFP) at 42h post-invasion. The results are the average of 4 independent experiments. Data are represented as mean ± SEM. See also figure 3D.

Figure S4. Related to Figure 4.

A- Integrative genomics viewer (IGV) plots for several PfAP2-I target genes. The log2 enrichment ChIP/Input for each ChIP replicate is shown. Grey blocks represent the peaks called by MACS2 (the peak number indicated above each peak), arrows represent genes, and small black, red and blue lines represent the presence of a predicted DNA motif associated with PfAP2-I-D3 or –D2 or –D1, respectively.

B- Within 250bp of the summit of each ChIP-seq peak, the DNA sequence motifs associated with PfAP2-I-D3 and –D2 are identified in the majority of peaks. However the motif associated with D1 is less commonly found. Black boxes represent the presence of the DNA motif within the peak and each row represents a PfAP2-I ChIP-seq peak. Although the D2 DNA motif can be seen in approximately 80% of the peaks containing a D3 DNA motif, the same trend is observed genome-wide. In a random test of 1000 genomic intervals similar in length to the trimmed peaks found by ChIP-seq, D2 and D3 motifs co-occur more frequently than individual occurrences (data not shown). Thus, the high percentage of motif co-occurrence seen in our data does not suggest biological significance. Black dots indicate genes tested by ChIP-qPCR. See also Figure 4.

Figure S5. Related to Figures 4 and 5.

A- Protein alignment of the three AP2 domains of PfAP2-I (PF3D7_1007700) and the AP2 domain of PfAP2-Sp (PF3D7_1466400). The conserved residues are highlighted in red. Highlighted with * are the mutated residues. See also figure 4A.

B- Western blots with anti-GFP indicate that PfAP2-I-GFP is expressed in the PfAP2-I-GFP-D1mut and PfAP2-I-GFP-D2mut strains. See also figures 4A-B.

C- Western blot probed with anti-GST antibody showing expression of GST-D1wt/mut, GST-D2wt/mut and GST-D3wt/mut. ni: non-induced, i: induced. See also figure 4C.

D- Sanger sequencing shows that the correct mutations were introduced into the coding sequence of D1 and D2 in PfAP2-I-GFP-D1mut or PfAP2-I-GFP-D2mut parasites (bottom), as compared to the parental strain, PfAP2-I-GFP (top). See also figures 4A-B.

E- Coomassie blue stained SDS-PAGE gels showing purified GST-fusion proteins used for PBM analysis: PfAP2-I AT-hook fused to the second AP2 domain (GST-AT-D2) (left), the purified AT-hook domain alone (GST-AT) (middle) and the second PfAP2-I AP2 domain (GST-D2) (right) - lysate (L), flow-through (FT), wash (W) and elutions (E). See also figure 4C.

F- PBM experiments performed using purified GST-fused proteins or bacterial lysate from cells expressing GST-AT-D2, GST-AT or GST-D2 (as control), showing that GST-AT-D2 does not bind the same DNA motif as GST-D2 but rather it binds the same motif as GST-AT. The assay was performed twice on different microarray versions (see experimental procedures) with independent protein purifications. See Table S6 for PBM position weight matrices.

G- Parasites expressing the NLS sequence of PfSIP2 (Flueck *et al*., 2010) fused to GFP were used as one of the IP controls. WB using anti-GFP antibodies shows that the NLS targets the protein to the nucleus but not exclusively. Histone H3 was used as a nuclear marker. See also figure 5A.

Figure S6. Related to Figure 5.

A- ChIP-qPCR of the same genes as shown in figure 2C demonstrates that PfBDP1-HA and PfCHD1-GFP bind the same genes as PfAP2-I-GFP. The results are shown as fold enrichment of ChIP performed with anti-GFP antibody versus ChIP performed with IgG and are the average of 3 biological replicates. Data are represented as mean ± SD. 3D7 *wt* parasites were used as a negative control (see also Figure 5). Figure 5C and S6B show some of the same data. *gexp02* (PF3D7_1102500) was included as an additional negative control for PfBDP1 binding due to the unexpected binding of the protein to *ama1*, a non PfBDP1 ChIP-seq gene target (PfBDP1 binding to this gene, as diagnosed by ChIP-qPCR, had been previously detected by Josling *et al*., 2015).

B- DNA motif discovery analysis using DREME shows that GTGCA is the most common DNA motif bound by PfBDP1-HA (1.5E-12) by ChIP-seq (top left). This DNA motif is identical to the one bound by PfAP2-I by ChIP-seq using DREME (bottom left). We also find a second DNA motif (C/GCACCT) bound by PfBDP1 (4.2E-5) (top right). This second DNA motif is similar to the one previously predicted using Gene Enrichment

Motif Searching (GEMS) to be enriched in micronemal gene promoters (Young *et al*., 2008) (bottom right). See also figure 5B.

C- Alignment of the nucleotide sequences found in the promoters of several micronemal and rhoptry neck genes shows that the second PfBDP1-bound motif (TAACT) (red) identified in C is conserved. Black dots highlight PfBDP1-HA ChIP-seq targets. We note that some of these genes are not PfAP2-I-GFP ChIP-seq targets.

D- The M20.1 ACAACT "micronemal" DNA motif is also found in the promoters shown in D. These results suggest that a TF responsible for regulating micronemal genes may bind a TAACT or ACAACT DNA motif, as both motifs are present upstream of these genes. However, an interaction between the TF and PfBDP1 may not be required for all micronemal genes to be transcribed, since not all of the genes were identified in the Josling *et al*. PfBDP1-HA ChIP-seq data (Josling *et al*., 2015).

Figure S7. Related to Figure 7.

A- ChIP-qPCR positive control demonstrating that PfBDP1 associates with a known target locus at the trophozoite stage. This gene was selected from the PfBDP1 ChIP-seq trophozoite target genes list from (Josling *et al*., 2015). Data are represented as mean ± SD and n=2. See also figure 7A.

B- Strategy used to C-terminally tag PfBDP1 with HA and the destabilization domain DD in the PfAP2-I-GFP expressing parasite. The presence of a PCR product for the reactions performed with primers 5 and 7 only in transgenic line demonstrates that the locus has been correctly modified. The primers sequences are listed in Table S1. See also figures 7B-D.

Table S1. List of primers used in this work. Related to the STAR methods.

Table S2. ChIP-seq data and GO-term searches. Related to Figures 2 and 3.

Table containing: 1)-6) MACS2 peak calling results for PfAP2-I-GFP ChIP-seq with anti-GFP/IgG antibodies for replicates 1, 2 and 3, 7) the list of peaks present in 2 out of 3 and 3 out of 3 biological replicates (merged peaks), 8) the list of merged peals trimmed to just include the overlapping genomic regions between biological replicates (trimmed peaks), the gene annotations corresponding to each trimmed peak and the peaks overlapping between PfAP2-I-GFP (this study) and PfBDP1-HA (Josling et al., 2015) ChIP-seq, 9)-11) GO-term searches

and 12)-14) a list of DNA motif occurrences bound by PfAP2-I AP2 domains and PfBDP1 as determined by FIMO and DREME.

Table S3. DNA microarray data. Related to Figure 3.

Results of the two DNA microarrays experiments performed with PfAP2-I-GFP and PfAP2-I-GFP::*msp5*MUT parasites (for a scheme of the collected time points see Figure S3G). Pearson correlation shows that the same time points were collected for PfAP2-I-GFP and PfAP2-I-GFP::*msp5*MUT parasites and that there is a high correlation between the two independent timecourses. Rnits (Sangurdekar, 2014), which attributes a p-value to each gene depending on the changes in transcript level, was used to compare the changes in gene expression on mutant parasites versus the parental line. The transcript that changed the most (lower p-value) on both timecourses on mutant parasites versus wt was HsDHFR, as expected, since PfAP2-I-GFP::*msp5*MUT express this gene (they were selected with WR) but PfAP2-I-GFP parasites do not. In order to determine which genes changed consistently between the two independent timecourses, Rnits was run on both timecourses and the data compared. Transcript levels for 49 genes had lower p-values than *msp5* but 61% of these (30/49) were genes encoding for either ribosomal RNA or antigenic variants or spliceosomal RNA, which are known to significantly and consistently change expression values between experiments (Rovira-Graells et al., 2012) (see also Figure S3H).

Table S4. IP proteomics results. Related to Figure 5.

IP proteomics results, including control IP experiments and SAINT results.

Table S5. ChIP-seq target genes microarray data. Related to Figure 2 and 6.

Expression patterns of ChIP-Seq target genes as determined by previous microarray studies (Bozdech *et al*., 2003; Hu *et al*., 2010; Le Roch *et al*., 2003).

Table S6. Position weight matrices for all AP2 domains tested in this study. Related to Figure 4.

Figure S1

A

pfrap2	TTTATA**TGCA**TAT--GAAG**TGCA**AATGA
pbrap2	CATTCA**TGCA**TAC--AAAG**TGCA**AAGAT
pvrap2	GTGTTA**TGCA**TAT--CAAG**TGCA**AAGAG
pkrap2	TAAGTA**TGCA**TCTACGTG**TGCA**GTTGA
pfrhoph2	TTTATA**TGCA**TATATGATGTGTTTTAGGT
pbrhoph2	AAAAA**TGCA**TGCCCTAATTATGTATATG
pvrhoph2	GCCTG**TGCA**GGCGCGAGCGGTTGCCCG
pkrhoph2	ATTTCA**TGCA**CTCCCCTGCGCCCTCACG
pfclag9	AATATA**TGCA**AATATTATTAGTATAAAAT
pbclag9	AAATCA**TGCA**TACTATTTATGTATTGTTCA
pvclag9	AGTGAA**TGCA**TACCCAAAAAAGCAAAAA
pkclag9	TAATCT**TGCA**GTAAAATGAAAATTTGGTT
pfasp	TACTAA**TGCA**TTT--AAAG**TGCA**ATTAT
pbasp	TGAATA**TGCA**CTT--ATTG**TGCA**TTAAA
pvasp	GTTGTA**TGCA**GCG-GAAA**TGCA**TGTAT
pfmsp4	TGAACA**TGCA**AAAN$_{47}$TCA**TGCA**TCTTT
pbmsp4	GATTAGT**TGCA**AAAN$_{34}$TGT**TGCA**ATATT
pvmsp4	CTTTGTG**TGCA**GAN$_{29}$CTG**TGCA**TATTT
pkmsp4	CATCACT**TGCA**TTTAGAAA**TGCA**GAAT
pfgap45	TATTTG**TGCA**ATAA-AAG**TGCA**TAGAA
pbgap45	GTCTATA**TGCA**ATG---AAG**TGCA**CCACA
pvgap45	GTGTATA**TGCA**ATA---AAG**TGCA**CCCCA
pkgap45	GCCTTTC**TGCA**ATA---AAG**TGCA**TCAGA
pfgap50	AATATA**TGCA**TATAAATATTACACTTAAA
pbgap50	CATTATA**TGCA**TAATTACACACACAGTATT
pvgap50	GTCCACA**TGCA**----------**TGCA**TGGGA
pkgap50	TCCCCTA**TGCA**CATACTTGTATGCTCACCC

B

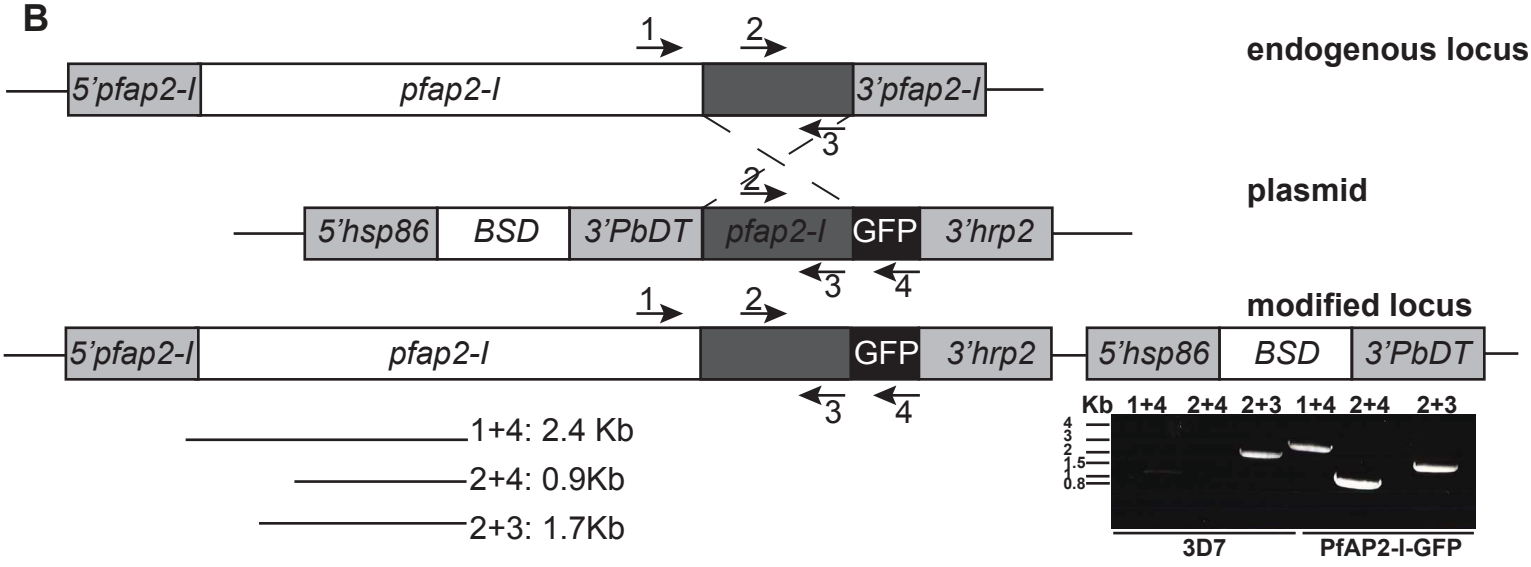


Figure S2

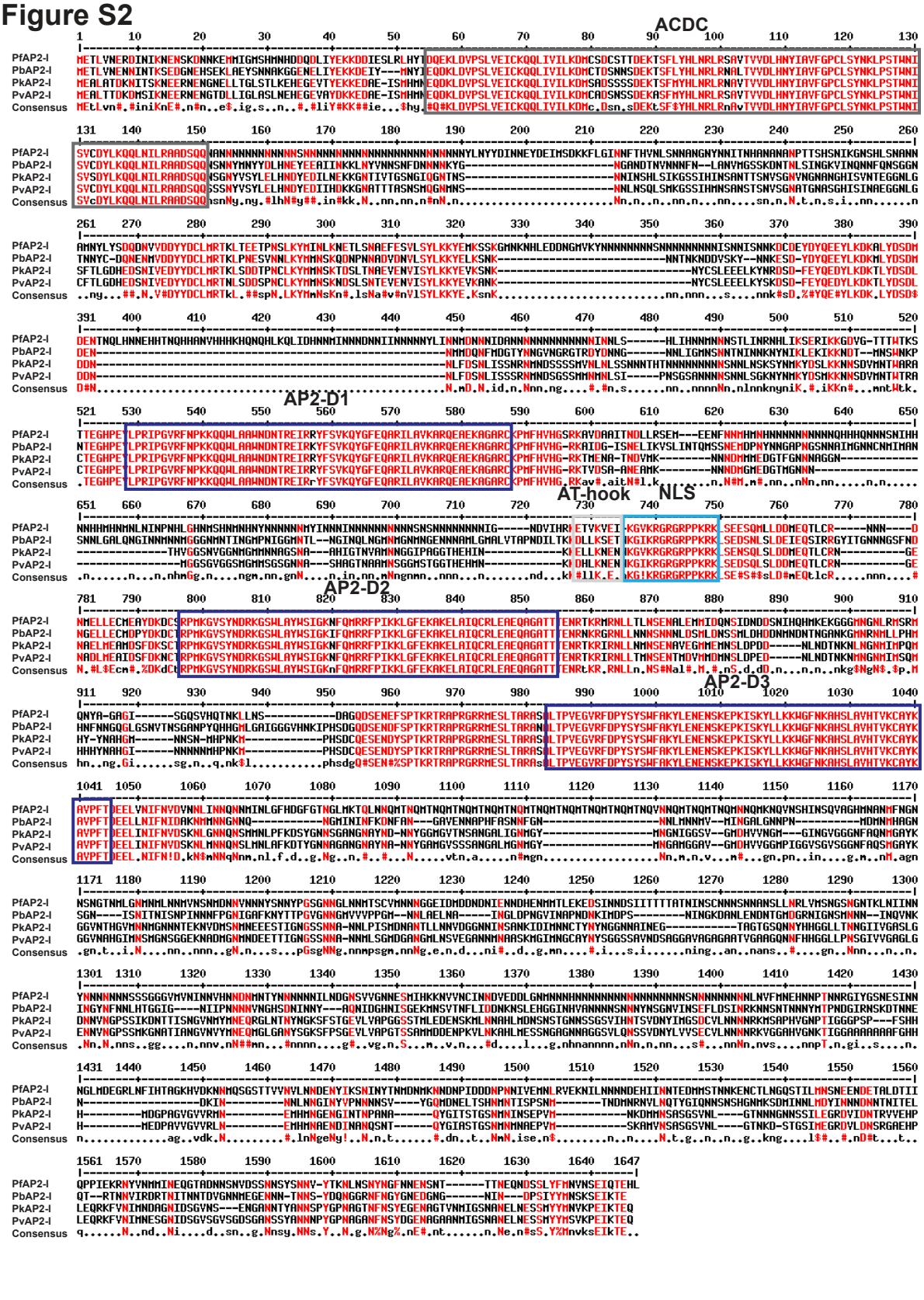

ACDC
AP2-D1
AT-hook
NLS
AP2-D2
AP2-D3
PfAP2-I
PbAP2-I
PkAP2-I
PvAP2-I
Consensus

Figure S3

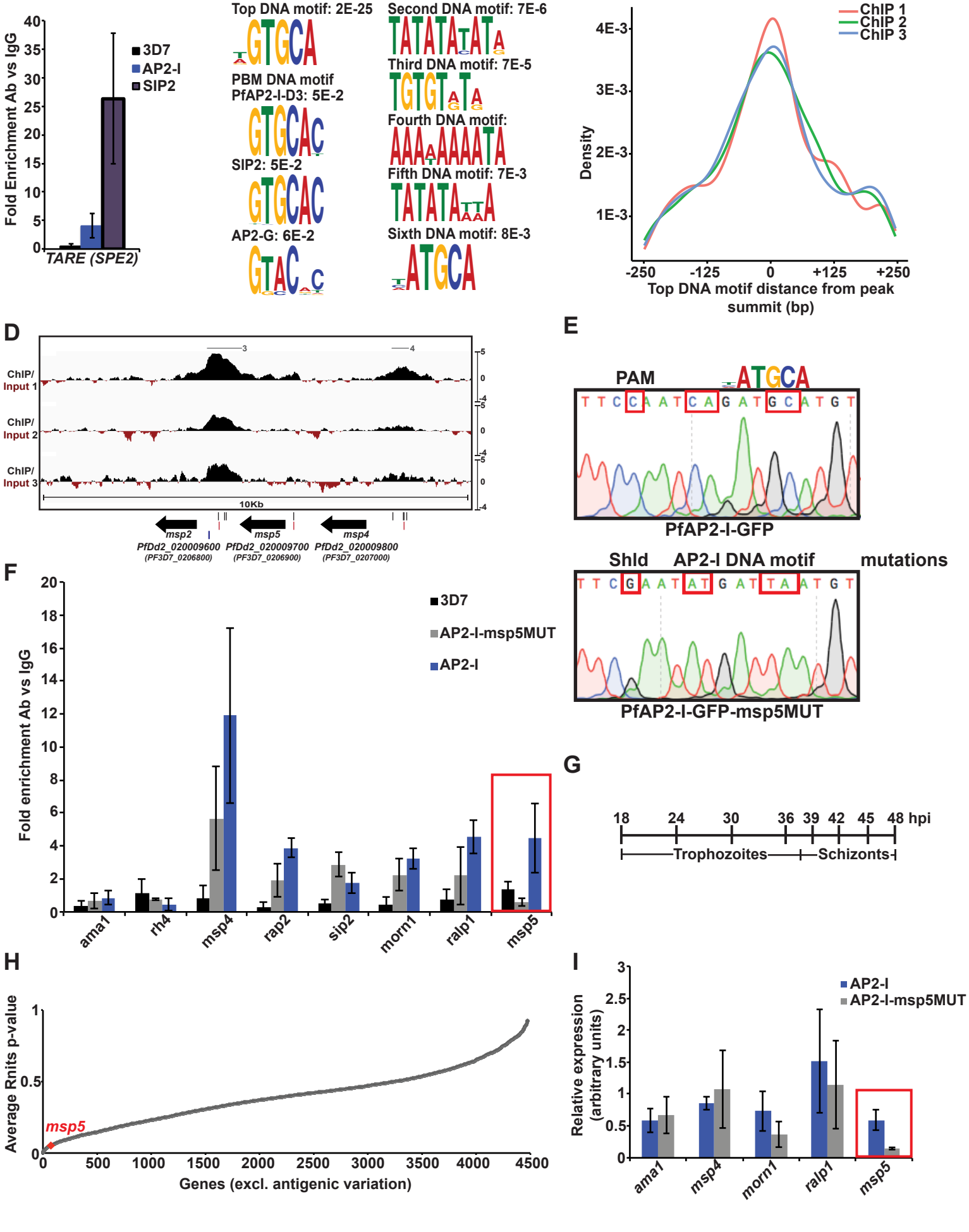

A
Fold Enrichment Ab vs IgG
3D7
AP2-I
SIP2
TARE (SPE2)
B
Top DNA motif: 2E-25
Second DNA motif: 7E-6
PBM DNA motif
PfAP2-I-D3: 5E-2
Third DNA motif: 7E-5
SIP2: 5E-2
Fourth DNA motif:
Fifth DNA motif: 7E-3
AP2-G: 6E-2
Sixth DNA motif: 8E-3
C
ChIP 1
ChIP 2
ChIP 3
Density
Top DNA motif distance from peak summit (bp)
D
ChIP/ Input 1
ChIP/ Input 2
ChIP/ Input 3
10Kb
msp2
msp5
msp4
PfDd2_020009600 (PF3D7_0206800)
PfDd2_020009700 (PF3D7_0206900)
PfDd2_020009800 (PF3D7_0207000)
E
PAM
PfAP2-I-GFP
ShId
AP2-I DNA motif
mutations
PfAP2-I-GFP-msp5MUT
F
Fold enrichment Ab vs IgG
3D7
AP2-I-msp5MUT
AP2-I
ama1
rh4
msp4
rap2
sip2
morn1
ralp1
msp5
G
18 24 30 36 39 42 45 48 hpi
Trophozoites
Schizonts
H
Average Rnits p-value
msp5
Genes (excl. antigenic variation)
I
Relative expression (arbitrary units)
AP2-I
AP2-I-msp5MUT
ama1
msp4
morn1
ralp1
msp5

Figure S4

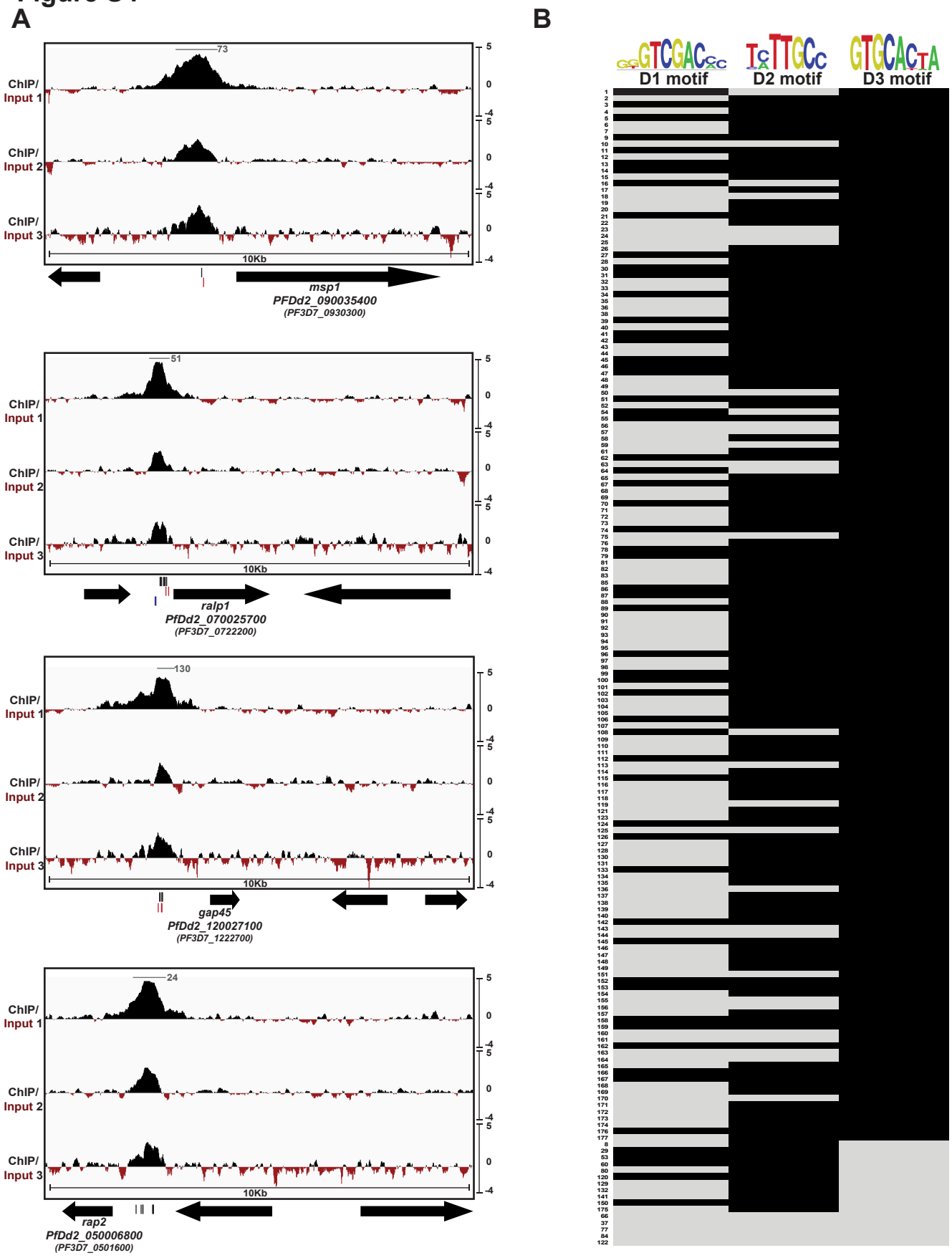

A
B
ChIP/ Input 1
ChIP/ Input 2
ChIP/ Input 3
10Kb
msp1
PFDd2_090035400
(PF3D7_0930300)
ralp1
PfDd2_070025700
(PF3D7_0722200)
gap45
PfDd2_120027100
(PF3D7_1222700)
rap2
PfDd2_050006800
(PF3D7_0501600)
D1 motif
D2 motif
D3 motif

Figure S5

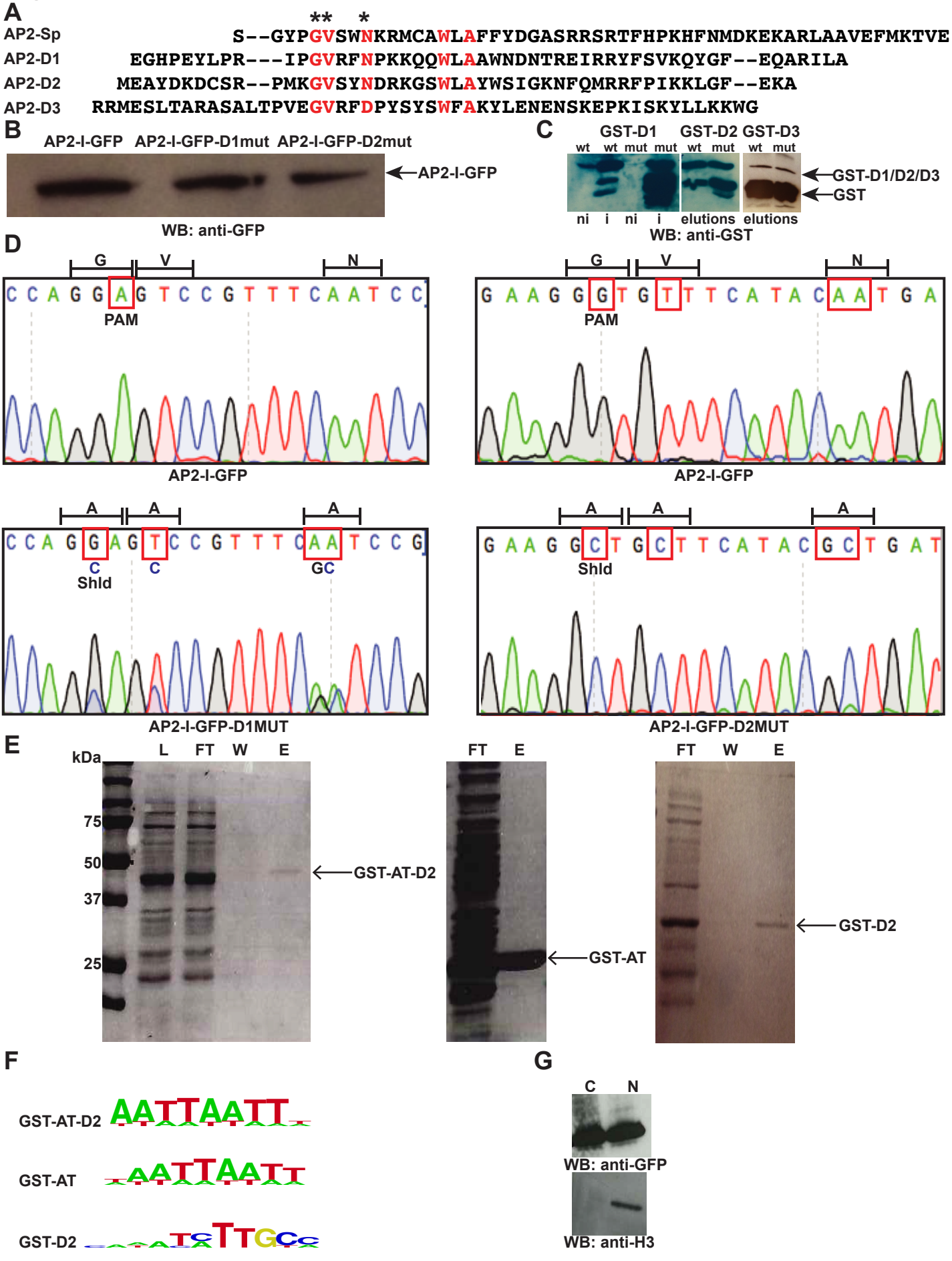
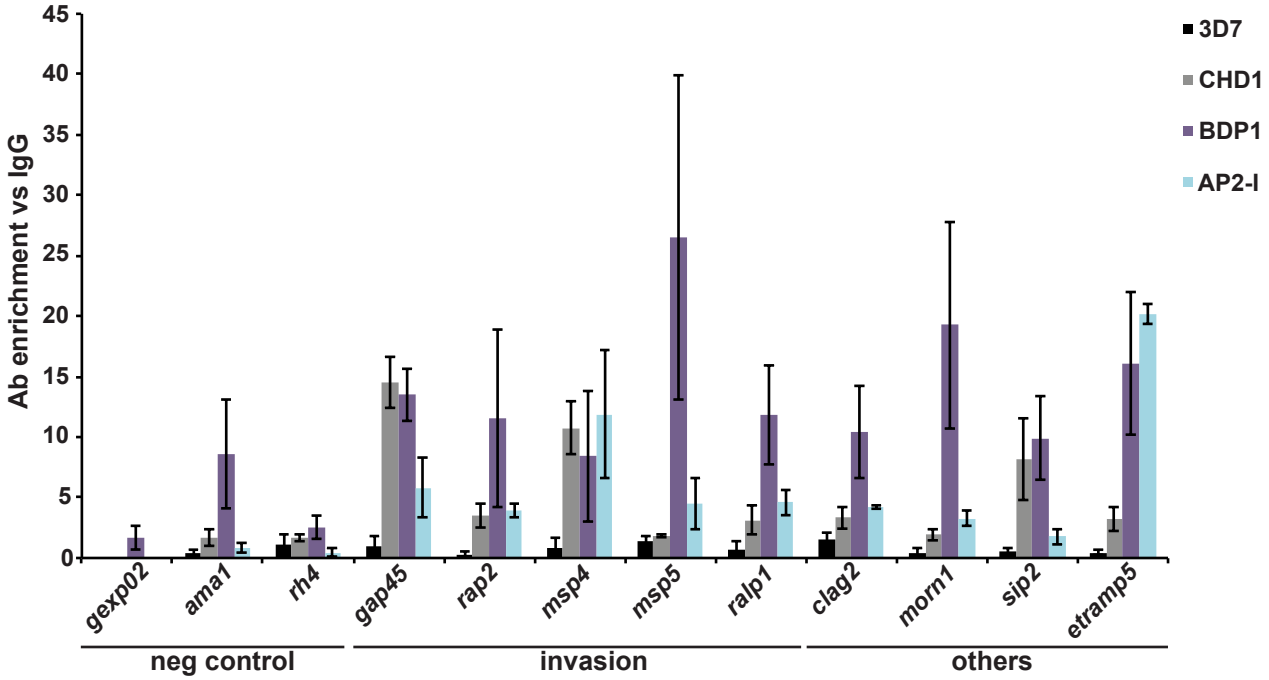
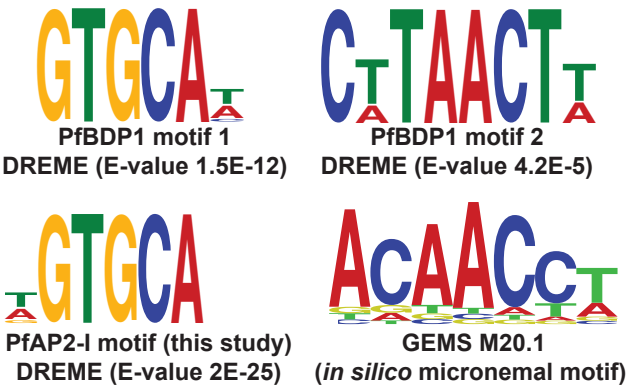

A
AP2-Sp S--GYPGVSWNKRMCAWLAFFYDGASRRSRTFHPKHFNMDKEKARLAAVEFMKTVE
AP2-D1 EGHPEYLPR---IPGVRFNPKKQQWLAAWNDNTREIRRYFSVKQYGF--EQARILA
AP2-D2 MEAYDKDCSR--PMKGVSYNDRKGSWLAYWSIGKNFQMRRFPIKKLGF--EKA
AP2-D3 RRMESLTARASALTPVEGVRFDPYSYSWFAKYLENENSKEPKISKYLLKKWG
B
AP2-I-GFP
AP2-I-GFP-D1mut
AP2-I-GFP-D2mut
WB: anti-GFP
C
GST-D1
GST-D2
GST-D3
wt wt mut mut wt mut wt mut
ni i ni i elutions elutions
WB: anti-GST
GST-D1/D2/D3
GST
D
PAM
AP2-I-GFP
AP2-I-GFP
AP2-I-GFP-D1MUT
AP2-I-GFP-D2MUT
ShId
E
kDa
L FT W E
75
50
37
25
GST-AT-D2
FT E
GST-AT
FT W E
GST-D2
F
GST-AT-D2
GST-AT
GST-D2
G
C N
WB: anti-GFP
WB: anti-H3

Figure S6

A



B



C

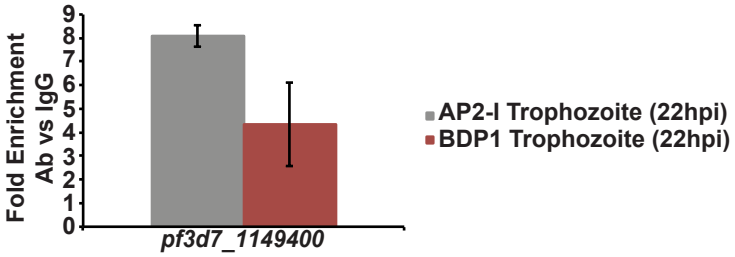
ama1 ●	ATTCGATAATTA**TAACT**AATAAATGTATTAAAA
eba181	TAATATGTATGA**TAACT**TTATTTAAAATTATAAA
mtrap	TTATATTTATTAT**TAACT**TTGTTTTGTTTTCCATGT
myoA	TTATATTTCCTCT**TAACT**TTACATGCTTAAATTAA
gap40	TTCTGTAAAATTT**TAACT**TGTCCACCCAAGAAAA
imc1a ●	TACTATAATTTTT**TAACT**TGGAAAAAAGGAAAAA
rh2a	AATTATTTTATAA**TAACT**GAA**TAACT**ATTTTTT
rh4 ●	TCTTTATATATAA**TAACT**TTGTAATTCCATAAAGA
ron3	ATACCAAATGTA**TAACT**AATATCATCAACAAGGA

D

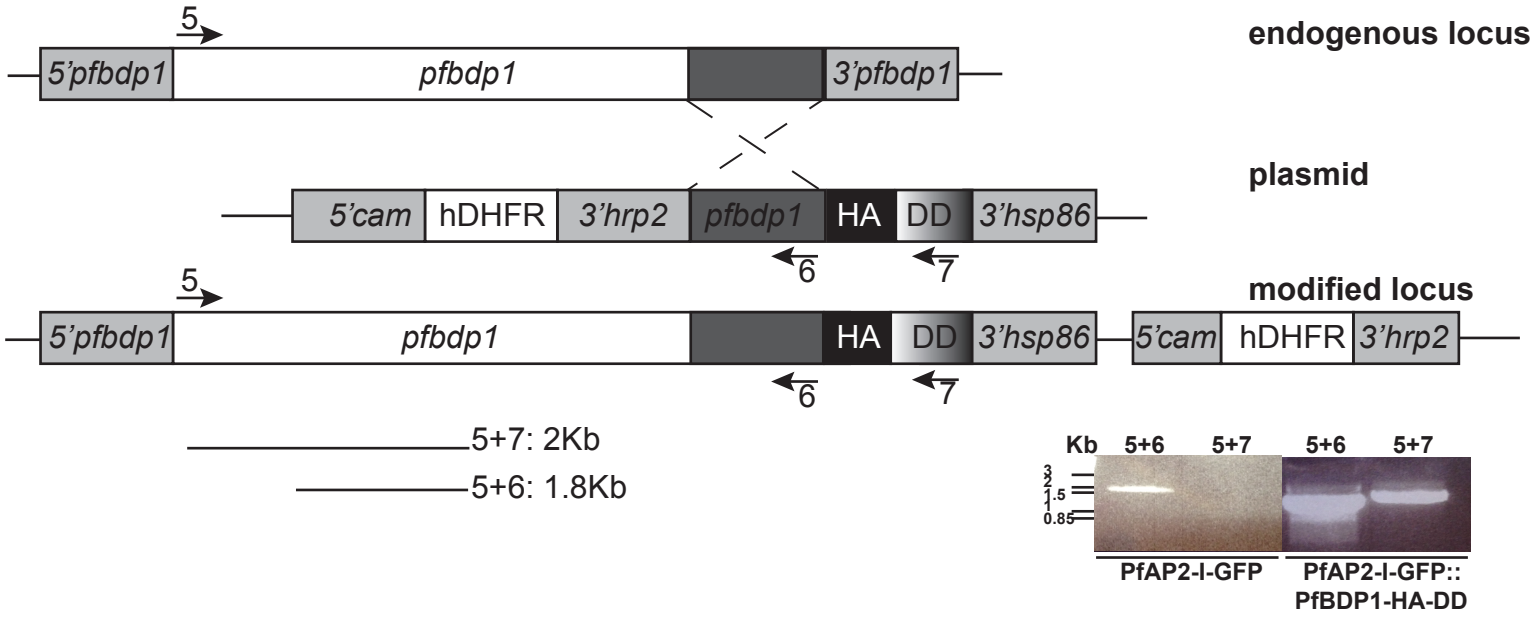
ama1 ●	CTGTTTTCTCTA**ACAACT**TTTTTTTTTTTTTTTT
eba181	ATAGTATACAT**ACAACT**TAATTAAAAAATAA
mtrap	ATAATAATAAA**ACAACT**ATTCTTTTATTTTAT
myoA	TTTTTTTTTTTTA**ACAACT**AAATTAGAAGTGTG
gap40	TGGATACATGA**ACAACA**CTTTTACTTTTTCATT
imc1a ●	AAATATAATTA**ACAACT**TTGTATTAATTCATG
rh2a	TGTAAATAATA**ACAACT**ATTCGGAATTAAAC
rh4 ●	CACTATGTTATT**ACAACC**TCATCTTTATGGAAT
ron3	CATCTGAATTGA**ACAACG**TTGTCTATATGTAAT

Figure S7

A


Fold Enrichment Ab vs IgG
9
8
7
6
5
4
3
2
1
0
pf3d7_1149400
AP2-I Trophozoite (22hpi)
BDP1 Trophozoite (22hpi)

B


5
5'pfbdp1
pfbdp1
3'pfbdp1
endogenous locus
5'cam
hDHFR
3'hrp2
pfbdp1
HA
DD
3'hsp86
plasmid
6
7
5
5'pfbdp1
pfbdp1
HA
DD
3'hsp86
5'cam
hDHFR
3'hrp2
modified locus
6
7
5+7: 2Kb
5+6: 1.8Kb

Kb
5+6
5+7
5+6
5+7
3
2
1.5
1
0.85
PfAP2-I-GFP
PfAP2-I-GFP::
PfBDP1-HA-DD