Uncovering the essential genes of the human malaria parasite *Plasmodium falciparum* by saturation mutagenesis


**INTRODUCTION:** Malaria remains a devastating global parasitic disease, with the majority of malaria deaths caused by the highly virulent *Plasmodium falciparum*. The extreme AT-bias of the *P. falciparum* genome has hampered genetic studies through targeted approaches such as homologous recombination or CRISPR-Cas9, and only a few hundred *P. falciparum* mutants have been experimentally generated in the past decades. In this study, we have used high-throughput piggyBac transposon insertion mutagenesis and quantitative insertion site sequencing (QIseq) to reach saturation-level mutagenesis of this parasite.

**RATIONALE:** Our study exploits the AT-richness of the *P. falciparum* genome, which provides numerous piggyBac transposon insertion targets within both gene coding and non-coding flanking sequences, to generate more than 38,000 *P. falciparum* mutants. At this level of mutagenesis, we could distinguish essential genes as nonmutable and dispensable genes as mutable. Subsequently, we identified 2680 genes essential for in vitro asexual blood-stage growth.

**RESULTS:** We calculated mutagenesis index scores (MISs) and mutagenesis fitness scores (MFSs) in order to functionally define the relative fitness cost of disruption for 5399 genes. A competitive growth phenotype screen confirmed that MIS and MFS were predictive of essential genes. Genes predicted to be essential included genes implicated in drug resistance—such as the “K13” Kelch propeller, *mdr*, and *dhfr-ts—as well as targets considered to be high value for drug development, such as *pck* and *cdpk5*. The screen revealed essential genes that are specific to human *Plasmodium* parasites but absent from rodent-infected species, such as lipid metabolic genes that may be crucial to transmission commitment in human infections. MIS and MFS profiling provides a clear ranking of the relative essentiality of gene ontology (GO) functions in *P. falciparum*. GO pathways associated with translation, RNA metabolism, and cell cycle control are more essential, whereas genes associated with protein phosphorylation, virulence factors, and transcription are more likely to be dispensable. Last, we confirm that the proteasome-degradation pathway is a high-value druggable target on the basis of its high ratio of essential to dispensable genes, and by functionally confirming its link to the mode of action of artemisinin, the current front-line antimalarial.

**CONCLUSION:** Saturation-scale mutagenesis allows prioritization of intervention targets in the genome of the most important cause of malaria. The identification of more than 2680 essential genes, including ~1000 *Plasmodium*-conserved essential genes, will be valuable for antimalarial therapeutic research.

---

**Saturation-scale mutagenesis of *P. falciparum* reveals genes essential and dispensable for asexual blood-stage development.**

(Top) A high-resolution map of a ~50-kb region of chromosome 13 depicts an essential gene cluster, including K13, that lacks insertions in the coding DNA sequence (CDS) but is flanked by dispensable genes with multiple CDS-disrupting insertions. (Left) The MIS rates the potential mutability of *P. falciparum* genes based on the number of recovered CDS insertions relative to the potential number that could be recovered through large-scale mutagenesis. (Right) The MFS rates the relative fitness of *P. falciparum* genes based on QIseq scores of transposon insertion sites in each gene.
Malaria caused by Plasmodium falciparum remains an insidious global health problem, with hundreds of thousands of deaths each year. Recently, there have been substantial reductions in disease intensity, in part through concerted recent use of artemisinin-combination therapies, but these gains are now threatened by emerging artemisinin-based combination therapy (ACT) treatment failures spreading across South East Asia (1, 2). If ACT resistance reaches Africa, a devastating rebound of disease is expected, as occurred with chloroquine resistance in the 1970s. The development of new antimalarial therapies, and identification and prioritization of new targets, is a priority. More than a decade after the completion of the P. falciparum genome, a substantial fraction of its genome still lacks functional annotation (3). Although CRISPR-Cas9 and other targeted endonucleases will accelerate functional genomic studies (4–6), their usefulness for genome-scale applications is restricted by the absence of nonhomologous end joining in Plasmodium parasites and by the extreme AT-richness of the P. falciparum genome, which reduces guide RNA target site abundance. Therefore, a critical need remains for large-scale genetic analysis to systematically identify essential genes and prioritize parasite metabolic pathways for drug discovery (7).

Large-scale genetic screening methods in model organisms rely on efficient scalable methods for genome engineering. Transposon mutagenesis using the piggyBac transposon, which preferentially inserts at the tetranucleotide target sequence TTAA, has been used to carry out whole-genome loss-of-function screens in many organisms (8–11). The highly skewed nucleotide composition of the P. falciparum genome, with >81% AT content, is an advantage for the application of piggyBac mutagenesis. The skewed composition results in a high density of TTAA sites, averaging one site per 70 base pairs (bp) over both coding and noncoding regions, in the genome loss-of-function screens in many organisms (8–11). The highly skewed nucleotide composition of the P. falciparum genome, with >81% AT content, is an advantage for the application of piggyBac mutagenesis. The skewed composition results in a high density of TTAA sites, averaging one site per 70 base pairs (bp) over both coding and noncoding regions, in the genome loss-of-function screens in many organisms (8–11). The highly skewed nucleotide composition of the P. falciparum genome, with >81% AT content, is an advantage for the application of piggyBac mutagenesis. The skewed composition results in a high density of TTAA sites, averaging one site per 70 base pairs (bp) over both coding and noncoding regions, in the genome loss-of-function screens in many organisms (8–11). The highly skewed nucleotide composition of the P. falciparum genome, with >81% AT content, is an advantage for the application of piggyBac mutagenesis. The skewed composition results in a high density of TTAA sites, averaging one site per 70 base pairs (bp) over both coding and noncoding regions, in the genome loss-of-function screens in many organisms (8–11). The highly skewed nucleotide composition of the P. falciparum genome, with >81% AT content, is an advantage for the application of piggyBac mutagenesis. The skewed composition results in a high density of TTAA sites, averaging one site per 70 base pairs (bp) over both coding and noncoding regions, in the genome loss-of-function screens in many organisms (8–11).
Defining gene dispensability by using mutagenesis index scores

Although there was a bias against insertions in CDSs, 2042 genes were disrupted by at least one piggyBac insertion (~38% of genes in the *P. falciparum* genome) (Fig. 1F). The remainder, 3357 genes (~62% of the genome) had no insertions in their CDSs, and some of these were also completely devoid of insertions in the surrounding intergenic regions (2.9% of genes) (Fig. 1C, fig. S4A, and table S4). In the 2042 mutable genes, insertion sites were distributed uniformly along the gene body CDSs, indicating that all disruptions have equivalent consequences for the disrupted gene (Fig. 1E). The uniform distribution

---

**Fig. 1.** A genome-wide saturation mutagenesis screen for *Plasmodium falciparum*. (A) Chromosomal map displays 38,173 piggyBac insertion sites from all mutants evenly distributed throughout the genome. (B) High-resolution map of a ~50-kb region of chromosome 13 depicts an essential gene cluster, including K13, flanked by dispensable genes with multiple CDS-disrupting insertions. (C) High-resolution map of a ~20-kb region without insertions includes three conserved genes of unknown function (PF3D7_1232700, PF3D7_1232800, and PF3D7_1232900) and a putative nucleotidyltransferase (PF3D7_1232600) (fig. S5). (D) A plot of all piggyBac insertions revealed that significantly fewer insertions were recovered from exon-intron regions compared with the proportion of available TTAA sites (fig. S1D) ($P < 2.2 \times 10^{-16}$, Fisher’s exact test). (E) Density of piggyBac insertion-site distribution revealed 75% fewer insertions recovered in transcriptional regions (blue) than intergenic 5’ (yellow) and 3’ (green) regions, depicted as relative distance upstream and downstream to a gene, respectively. (F) This study determined that under ideal culture conditions for asexual blood-stage growth, 38% of genes in the *P. falciparum* genome have mutable CDSs, whereas 62% of genes have nonmutable CDSs, which includes 12% with tentative classification.
of transposon insertions throughout the genome is independent from chromatin structure and gene locations (fig. S4B). We therefore reasoned that the recovery of one or more insertions within the CDS of disrupted genes was an indicator of gene dispensability for in vitro asexual blood-stage growth. Therefore, mutable genes are subsequently referred to as dispensable, whereas the absence of any insertions in the CDS could be considered an indicator that disruptions are lethal, and subsequently, these genes are referred to as essential. However, the essential classification of 677 genes without insertions, which are small or with lower TTAA, is tentative because they are below the average distance between recovered insertions (<613 nucleotides [nt]) (table S5). To quantify the evidence for dispensability or essentiality of each gene, we developed a mutagenesis index score (MIS) based on the number of identified piggyBac insertions relative to the number of available TTAA target sites within that gene (Fig. 2A and table S5). Genes with a higher MIS (on a scale of 0 to 1) were considered to have a higher possibility of dispensability, whereas those with a lower MIS were considered to have a higher possibility of essentiality. The 2000 genes with highest and lowest MISs represent mutability characterization with the strongest confidence.

**Biological processes identified as dispensable and essential**

Mutants in the lowest MIS quartile were enriched in core metabolic processes shared across eukaryotes, which would be predicted to be essential. Mutants in the highest MIS quartile were enriched in parasite-specific multigene families that interact directly with the host and that are known to be partially functionally overlapping and contain many redundant genes (Fig. 2B). This screen was carried out by using in vitro cultured parasites, and many of these in vitro dispensable genes, such as those for antigenic variation and cytoadherence, have greater importance in vivo in human infections. The dispensability of parasite processes involved in host-parasite interactions contrasted with the essentiality of internal parasite processes, such as those associated with RNA metabolism (Fig. 2, C and D; fig. S6; and table S6). RNase II, a gene implicated in a posttranscriptional regulatory mechanism relevant to severe malaria, is an example with a low MIS, which is consistent with previous reports of the locus being refractory to disruption (20). Other genes involved in general RNA metabolism that have low MISs include both widely conserved RNA-binding proteins (such as PABP) and apicomplexan-specific RNA-binding proteins likely to have more specific, mostly unidentified, regulatory targets. The *P. falciparum* genome has an abundance of RNA-binding proteins that are largely functionally uncharacterized, and our analysis suggests that many of these genes are likely to be essential.

MIS therefore correlates with what is known broadly about the importance or redundancy of metabolic pathways. To further validate that MIS is a good predictor of essentiality, we analyzed genes that have been the focus of drug or vaccine development. Genes strongly predicted to be essential based on MIS included the “K39” Kelch propeller implicated in artemisinin resistance, as well as other genes implicated in drug resistance, such as DHFR-TS, MDR, and AAC2, which are involved in pyrimethamine, mefloquine and atovaquone resistance, respectively. Other genes classified as essential based on MIS included ones considered high-priority blood-stage drug targets such as PFKG and CPK65 (fig. S5). By contrast, most blood-stage vaccine candidates were dispensable, with the notable exception of RH5.

Not unexpectedly, sporozoite vaccine candidates CSP and TRAP, which are essential for sporozoite development in mosquitoes but are not required in blood-stage development (21), had high MIS. By contrast, the gene for pore-forming protein cell traversal of ookinete and sporozoites (CetTOS) had a low MIS. CetTOS is important for these parasite migratory phases and is emerging as a pre-erythrocytic-transmission-blocking vaccine candidate; its low MIS indicates it may also have an essential function in blood-stage infections, making it a potential multi-stage vaccine target. MIS analysis also revealed likely essential genes that are specific to human *Plasmodium* parasites but absent from rodent-infesting species, such as the lipid metabolic genes (PCD and PMT) that are crucial in the development of mosquito-transmissible sexual stages in *P. falciparum* (22, 23).

**Mutant growth fitness as a measure of gene dispensability**

We previously developed a phenotyping method in which a pool of piggyBac mutants were grown as mixed-mutant pools over multiple generations, and the number of reads for each piggyBac insertion site was quantified by using next-gen sequencing in order to measure parasite growth rates (17). We adapted this method to generate a second quantitative measure of gene dispensability independent from MIS, by comparing the normalized number of reads from each insertion site to the total pool of reads across all *P. falciparum* mutants in the saturation mutagenesis screen (Fig. 2E and table S5). This mutagenesis fitness score (MFS) serves as a proxy for mutant growth fitness. MFSs strongly correlated with MISs, despite the fact that they are independent of each other (Fig. 2F). The MFS of mutable genes was high, in keeping with the MIS prediction of dispensability, whereas predicted essential nonmutable genes had a low MFS (Fig. 2G). Intermediate MISs and MFSs were weakly correlated, in keeping with the lower confidence that we can place on the essentiality or dispensability of genes with these intermediate scores.

We used a competitive in vitro growth screen (17) to phenotype four separate pools of mixed-mutant populations so as to provide additional validation of the correlation between MIS and a gene’s importance for growth fitness. Mutatable genes with high MISs and low or minimal fitness cost were competitive growth “winners,” whereas growth fitness “losers” had relatively low MISs (Fig. 3A and table S7). Relatively few mutations had little or no fitness cost, as measured with MFS, resulting in a disproportionate number of mutant losers (Fig. 3B). Overall, both the MIS and MFS were predictive of the fitness cost for in vitro asexual growth (Fig. 3C), and between these metrics, we were able to predict separate 5399 genes in the *P. falciparum* genome into nonmutable and mutable categories (Fig. 3, D and E).

**Association of essentiality with genome structure and transcription patterns**

Dispensable and essential genes rarely occurred as single isolated genes but rather occurred as multigene clusters reminiscent of conserved syntenic blocks (Fig. 1, A and C, and fig. S4A) (24). We therefore assessed the relationship between evolutionary conservation of genome structure and essentiality or dispensability by using previously defined syntenic relationships between *Plasmodium* spp. (25). Syntetic genes indeed had lower MISs, suggesting conserved essential functions across *Plasmodium* spp.; conversely, nonsyntetic genes were more likely to be dispensable (Fig. 4A). Mapping chromosomal regions with synteny breaks, which typically harbor gene duplications and paralogs, showed patterns of clustering of essential and dispensable genes in syntenic and nonsyntenic chromosomal regions, as seen in examples for chromosomes 13 and 10 (Fig. 4, B to E).

Transcription metrics for each gene based on maximum FPKM (fragments per kilobase of exon per million fragments mapped) values (26) were examined for correlations with essentiality and dispensability (Fig. 5A). This analysis showed that essential genes are expressed at significantly (P = 2.2 × 10⁻⁵) higher levels throughout both the asexual and sexual life cycle phases, indicating that they may have critical functional roles throughout both the mosquito and human stages and therefore represent potential multistage drug targets (Figs. 5, B and C). Within the intraerythrocytic cycle, genes with peak expression during the trophozoite stage were more likely to be essential than those expressed during other stages (Fig. 5C). During the trophozoite stage, *P. falciparum* must acquire nutrients from the host and remodel the infected erythrocyte to avoid host defenses, such as clearance by the spleen. By contrast, dispensable processes were enriched at either end of intraerythrocytic development, reflecting parasite stages with a greater need for functional redundancy, such as those involved in host interaction (such as erythrocyte invasion and antigenic variation), and are hence more likely to undergo duplication or paralog evolution. Not unexpectedly, ~41% of the dispensable genes are predominantly expressed in sexual stages (Fig. 5C). These genes may be essential for transmission to or from the mosquito but do not appear to have an essential function for asexual blood-stage development. The high coverage of sexual-stage genes in this screen emphasizes the potential for
Fig. 2. Identification of dispensable and essential genes through MIS and MFS. (A) The MIS rates the potential mutability of *P. falciparum* genes based on the number of recovered CDS insertions relative to the potential number that could be recovered. Genes known as dispensable or essential are highlighted. (B) MIS violin plots of GO processes grouped from lowest to highest dispensability, according to gene functional annotations. (C) MIS plots and (D) high-resolution chromosome maps highlighting important genes of interest for RNA metabolism (~20 kb, +20 kb) (MIS plots of other genes of interest are provided in fig. S4). (E) The MFS estimates the relative growth fitness cost for mutating a gene based on its normalized Qseq sequencing reads distribution. (F) MIS has significant correlation to MFS (Pearson’s $R = 0.67$, $P < 2.2 \times 10^{-16}$ compared with permutation). (G) The first and second MFS quartiles were composed primarily of nonmutable genes, the fourth quartile was composed mostly of mutable genes, and the third quartile had nearly equal numbers of both.
piggyBac-based screens for identifying sexual phenotypes.

Evolutionary conservation of apicomplexan gene essentiality

We compared the data from our genome-wide saturation screen with data from large-scale genome-sequencing studies of *P. falciparum*, as well as data from recent large-scale mutagenesis studies of apicomplexan genomes (6, 18, 27–29), to evaluate the conservation of gene essentiality across the organisms’ lineages. Essential *P. falciparum* genes were more highly conserved across *Plasmodium* spp. (Fig. 5D and fig. S8A), are less likely to have a paralog (Fig. 5E and fig. S8B), and encode genes with less genetic variation among *P. falciparum* clinical isolates (Fig. 5F and fig. S8C). Of the *P. falciparum* genes studied here, 2083 and 1998 have orthologs in the more distantly related parasites *Toxoplasma gondii* (6) and *P. berghei* (29), respectively. Overall, there was a strong correlation in gene function between species, particularly with genes predicted as essential (Fig. 5, G and H, and fig. S8, D and E).

Genes that are dispensable in *P. falciparum* were also more likely to be dispensable in *P. berghei*. By contrast, the correlation between the dispensable...
as essential, whereas individual genes linked to the degradation processes were experimentally defined as being involved in ubiquitin-dependent degradation pathways (Fig. 6, A to C; fig. S9, A to C; and table S9) with essentiality of potential druggable targets and pathways. Comparing the relative number of essential, versus dispensable, genes within specific GO biological processes, molecular functions, and cellular components provided a ranking of the essentiality of potential druggable targets and pathways (Fig. 6, A to C; fig. S9, A to C; and table S8). For example, nearly all genes annotated as being involved in ubiquitin-dependent degradation processes were experimentally defined as essential, whereas individual genes linked to microtubule motility and antimicrobial activity are experimentally defined as dispensable based on MIS. RNA metabolism and translation-related processes are also highly essential, supporting emerging evidence for the importance of posttranscriptional and translational control (30).

**Essentiality of the proteasome-degradation pathway**

Recent genetic analysis of artemisinin-combination therapy resistance has linked drug resistance to cellular stress-response mechanisms involving the *P. falciparum* ubiquitin/proteasome system (31, 32). Genes of the proteasome-degradation pathway were well represented in the piggyBac mutagenesis screen, and 54 of the 72 genes could be classified as essential genes according to our data (fig. S10A), further strengthening the priority of the proteasome degradation pathway as a high-value druggable target (33). We validated this link between artemisinin sensitivity and proteasome inhibition sensitivity using a set of single-insertion piggyBac mutants previously defined with chemogenomic profiling to be part of an artemisinin-sensitivity cluster, including a mutant of the K13 Kelch propeller gene (16, 34). We found 10-fold increased sensitivity to the proteasome inhibitor Bortezomib in mutants of the artemisinin-sensitivity cluster, which is significantly different relative to wild-type NF54 or mutants with distinct chemogenomic profiles (fig. S10, B to E), providing additional support for the association between the ART (Artemisinin) mechanism of action and proteasome function.

**Summary**

Our genome-wide saturation mutagenesis screen in the major human pathogen *P. falciparum* has defined genes essential for parasite survival during the blood stage and provided critical functional data for prioritizing high-value drug targets and pathways. The complete characterization of the essential genome with high-throughput saturation mutagenesis will help open new frontiers for antimalarial therapeutic research. The methodology also opens the way for new systematic functional screens for other phenotypes, such as transmission and cytoadherence.

**Materials and Methods**

**Parasite culture**

All *P. falciparum* parasite NF54 and piggyBac mutants were cultured in complete RPMI 1640 with 5% hematocrit (medium containing 0.5% Albumax II, 0.25% sodium bicarbonate and 0.01 mg/ml gentamicin). All parasite cultures were maintained by standard methods (35).

**piggyBac Transfection**

High-efficiency transfection (96-well plate method) was carried out on NF54 schizonts purified by magnetic column (MiltyenyiBiotec, CS column) using a transposon plasmid (pLBacII-HDH, containing selection marker human dhfr) and a transposase-expressing helper plasmid (pDCTH) (19, 35). RBCs were first loaded with plasmid DNA by electroporation using Gene PulserXCell+CE Module (BioRad). A total of 10 million schizonts, 1200 μg plasmid DNA and 600 μg helper plasmid DNA were used per 96-well plate to achieve maximal transfection efficiency. The drug WR99210 (final concentration 2.5 nM) was added to each plate for selecting transfected parasites, using a robot (Integra VIAFLO 96) to change media with drug every day for 5 days. Transfections were performed in batches over the course of 4 to 6 weeks. Each 96-well plate was cultured in triplicate and cryopreserved in duplicate. One hundred 96-well plates were used to generate the large pools for sequencing (fig. S1A).

**References**

For maximum recovery of unique mutants, the transfected culture was distributed into 96-well plates by robot (200 μl each well) and grown under two rounds of drug selection. Wells were screened for parasites by PCR and Giemsa-stained thin blood films. >50% of wells became parasite-positive within three weeks. Each positive well with ≥9 unique mutants (one well = one “Mixed Population,” or MP) was cryopreserved in duplicate (two 96 well plates). Additional 96-well plates were used to generate the pools that were then cultured in a T75 flask and harvested after two cycles growth by standard methods for genomic DNA isolation for QIseq (17). Although mutants grow at different rates, the pools can be expanded for at least 12 asexual life cycle generations (“cycles”) without significant loss in detectable diversity. Importantly, each mutant pool can be regenerated and cloned from their original MPs.

Datasets of reference genomes, transcriptome data and epigenetics data

The NF54 reference genome can be downloaded from (ftp://ftp.sanger.ac.uk/pub/project/pathogens/Plasmodium/falciparum/NF54/Assembly/V1_morphed). The QIseq sequencing followed previously published methods (17). Ortholog counts, paralog counts, and non-synonymous/synonymous (NonSyn/Syn) ratios can be downloaded from PlasmoDB (http://plasmodb.org/plasmo) (36). Previously published RNA-seq expression data (26, 37) are used in computational analysis. Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq) data of seven time points (0, 6, 12, 18, 24, 30 and 36 hour) were used to measure chromatin accessibility (38). The motif information of transcription factor ApiAP2 binding sites is available in previously published work (39). The genome synteny across five Plasmodium species is available in published work (25).

Identification of >38,000 mutants carrying piggyBac insertion

The QIseq sequencing generated a total of 3,301,112 raw reads. First, all the raw reads were filtered by the criteria of matching expected transposon target site “TTAA.” The motif information of transcription factor ApiAP2 binding sites is available in previously published work (38). This filtering step yielded 2,042,147 reads, which is >sixfold genome coverage. The insertion site QIseq signal was calculated as the reads count of the filtered QIseq reads. To compare the insertion signals...
between different QIseq runs, we normalized for each insertion \( i \) as

\[
S_i = \frac{c_i}{\sum_{j=k}^{R} c_j} \times 250,000
\]

where \( S_i \) represents the signal of the insertions \( i \) with reads counts \( c_i \) in run \( R \). All the piggyBac insertions with reads in both 5' and 3' ends were treated as true insertions by using an extremely high level of accuracy in target site (target motif TTAA ~ 99%). To identify true insertions with reads only in single 5' or 3' ends, we used a method that selected the most accurate signal cutoff based on the ratio of the number of input reads/ the number of output disrupted genes.

Fig. 6. Differentiating dispensable and essential genes and discovering high-priority druggable targets and pathways. (A) Functional annotations of biological processes are represented by the \( P \) value, and the \( x \) axis shows the fraction of the genes with MIS > 0.5. Each GO term is assigned a \( P \) value on the \( y \) axis to represent the tendency to be essential or dispensable. Essentiality is indicated on a spectrum of red (essential) to blue (dispensable) and circle sizes indicate the GO term enrichment. (B and C) Boxplot of (B) molecular processes and (C) cellular components shows the MIS distribution generated by 1000× sampling of the number of genes in the query GO-term category. Left (red) and right (blue) triangles indicate GO terms with significantly lower or higher MIS (\( P < 0.05 \) compared with background), respectively; the heatmap represents the essentiality defined as the fraction of genes per GO term with MIS > 0.5.
We computed the number of gene CDSs targeted by the insertions upper cutoff (i.e., more accurate). We noticed that there was a single large and sharp increase in the number of CDSs targeted by the insertions with reads number increasing from cutoff 2.0 to cutoff 2.3 (P < 0.05 compared with changing another cutoff) (fig. S2E). The result indicated a large amount of false positive may be included when setting cutoff ≥ 2.3. Therefore, the normalized reads count = 2.3 was set as the lower bound for identifying true insertions.

**Mutagenesis saturation computational validation**

To evaluate the levels of whole-genome mutagenesis and eventual saturation, we used a computational method based on sampling. First, a specific number of insertions were randomly extracted and the fraction of genome elements targeted were counted. This procedure was repeated 1000 times for each query insertion number in fig. S1C. A log-based curve was observed by plotting the mutational events and the median number of genome elements targeted by insertions. The plateauing of the curve at about 20,000 mutational events indicated that a very small number of new genome elements could be targeted even with very large amount of insertions. Importantly, the real data-based sampling curve was very similar to our independent mathematical predictions of Negative Binomial distributions prior to the start of the saturation mutagenesis experiments (fig. S1D). Together, both our mathematical model and real-data computational sampling show that >33,000 mutants represent saturation-level mutagenesis of the entire set of protein-encoding genes in the *P. falciparum* genome.

**Sampling methods for accounting for fragment length differences and nucleoside composition bias**

For genes in the *P. falciparum* genome, 5′ upstream and 3′ downstream intergenic regions displayed similar TTAA densities (median number of TTAA sites per 100 bp was 1.82 and 1.89, respectively). However, they have significantly different region length distributions (median values are 1681 bp and 926 bp in 5′ and 3′ direction, respectively; P < 0.001 Wilcoxon test). Therefore, there are different numbers of genomic TTAA sites in 5′ and 3′ intergenic regions, as an intrinsic feature of the *P. falciparum* genome. To account for this region-length and total-TTAA bias in our QiSeq calculations, the TTAA sites in 3′ intergenic regions were sampled based on the TTAA number in 5′ intergenic regions, so that our comparison was based on two identical background target-site distributions. Distribution of insertions and TTAA sites (Fig. 1E) were plotted based on these sampling data.

**MIS calculation**

To quantify the mutability of every protein-encoding gene in the *P. falciparum* genome, we first checked the number of *piggyBac* insertions located inside the gene CDS regions. To keep the criteria stringent, the insertions located on the very end (distance from insertion to TSS > 99% of the CDS length) of the CDSs were not considered, due to a significantly higher number of insertions at the last 1% of the CDS compared with other regions of the CDS (fig. S2F). We calculated the initial score MS of gene $MS_g$ based on the equation

$$ MS_g = \log \left( \frac{N_g + 1}{D_g} \right) $$

where $N_g$ represents the number of insertions on gene $g$. $D_g$ is the TTAA density of the gene $g$, which could be represented as the number of TTAA per kb of the CDS. We found that the $MS$ could be decomposed into two mixed Gaussian distributions. Therefore, a binary variable $\pi_g$ was used to model whether the gene $g$ is dispensable or not: $\pi_g = 1$ corresponds to dispensable and vice versa. The probability of observing a $MS_g$ of gene $g$ is mixture of two Gaussian distributions:

$$ P(MS_g) = \sum_{\pi_g=1} P(\pi_g)N(MS_g | \mu_{\pi_g}, \sigma_{\pi_g}) $$

$$ P(\pi_g = 1 | MS_g, \Theta) = \frac{\sum_{\pi_g=1} P(\pi_g)N(MS_g | \mu_{\pi_g}, \sigma_{\pi_g})}{\sum_{\pi_g=0} P(\pi_g)N(MS_g | \mu_{\pi_g}, \sigma_{\pi_g})} \quad (4) $$

where $\Theta$ is the parameter space. The MIS was defined as the posterior distribution to be dispensable.

**MFS calculation**

For the mutants that were subjected to competitive growth essay, we calculated MFS to represent the comparative growth fitness of a mutant. The QiSeq sequencing reads distribution reflects the fitness of unique *piggyBac* mutants in the competitive-growth assays. The start and the end of the growth assay could be represented as $t_0$ and $t$. At the end of the assay, $\bar{a}_{t_g}^m$ indicated the relative abundance of each mutant $m$, targeted into the CDS of gene $g$ in sample $v$. Therefore, the abundance of the mutant of the gene $g$ at the end of the assay could be represented as $\sum_{m} a_{t_g}^m$. The $\sum_{m} a_{t_g}^m$ was directly proportional to the QiSeq reads number $\sum_{m} b_{r_g}^m$. Here, $a_{t_g}^m$ represents the normalized reads number of mutant $m$ targeting gene $g$ on CDS, which is $\sum_{m} a_{t_g}^m = \sum_{m} a_{t_g}^m \cdot \sum_{m} b_{r_g}^m$. The QiSeq reads starting with ‘TTAA’ were used to measure $r_{t_g}^m$. To normalize different sample $v$, the average value of $\sum_{m} a_{t_g}^m$ was calculated as

$$ MS_g = \log \left( \sum_{m} b_{r_g}^m / \sum_{m} b_{r_g}^m \right) \quad (5) $$

Here, $I(\cdot)$ is the indicator function $\sum_{m} b_{r_g}^m / \sum_{m} b_{r_g}^m$ as the number of mutants targeting gene $g$ in all samples. At the start of the assay $t_0$, the relative abundance of each mutant equal to the background TTAA density $D_g$. Finally, the MFS is calculated as:

$$ MFS_g = \log \left[ \sum_{m} b_{r_g}^m / \sum_{m} b_{r_g}^m \right] \quad (6) $$

Therefore, the MFS calculation based on individual mutant abundance is a proxy for competitive growth fitness in in vitro blood stage.

**GO analysis**

All the Gene Ontology (GO) terms were downloaded from (http://plasmodb.org/plasmo/) (27). For each specific GO term $g$, the MIS distribution in the term $g$ could be represented as $MS_g$. The number of genes in term $g$ is $N_{tg}$. We asked whether the genes in the term $g$ are more prone to be dispensable or essential. To answer this question, we compared the median value of $MS_g$ with sampled data distribution as background. The genes with the same number of $N_{tg}$ were sampled out and represented as $S_g$. This procedure was repeated 1000 times and the median MIS of each $S_g$ ($t \in 1, 2, 3, ...1000$) was calculated as $MS_{S_g}$. The P value was calculated as

$$ P_g = \frac{\sum_{t=1}^{1000} I(MS_g < MS_{S_g})/1000}{MS_{S_g} < 0.5} \quad (7) $$

Here, $I(\cdot)$ is the indicator function.

**Half-maximal inhibitory concentration (IC50) estimation**

To calculate IC50, the dose-response data (e.g., drug concentrations $c_1, c_2, ...c_n$ and growth inhibition $q_1, q_2, ...q_n$) were used to fit a Hill Equation.

$$ q = U + \frac{U - D}{1 + 10^{(C_50 - \log(c))}} \quad (8) $$

where $c$ was drug concentrations in logarithmic form. The parameter $C_50$ was the estimate of IC50. The goodness of fit was calculated from dose-response data and the Hill Equation.

**REFERENCES AND NOTES**


17. I. F. Broner et al., Quantitative insertion-site sequencing (Qiseq) for high throughput phenotyping of transposon mutants. Genome Res. 25, 980–989 (2016).
29. S. S. Vembar, D. Droll, A. Scherf, Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. PLOS Pathog. 6, e1002135 (2010).
30. S. R. A., and R. H. J.; writing group consisted of M. Z., C. W., T. D. O., J. O.; computational analysis was performed by C. W., M. Z., T. D. O., J. O.; quantitative insertion-site sampling was performed by M. Z., I. F. B., M. M., and J. B.; computational analysis was performed by C. W., M. Z., T. D. O., J. O., S. R. A., and R. H. J.; writing group consisted of M. Z., C. W., T. D. O., J. O., C. R., R. H. J., and J. A.; and the study was conceived and directed by J. C. R., R. H. J., and J. A.; Competing interests: U.S. patent 7932088 (26 April 2011). “High efficiency transformation of Plasmodium falciparum by the Lepidopteran transposon, piggyBac;” is held by inventors J. H.A., M. J. Fraser Jr., B. Balu, and D. A. Shoue. The invention relates to use of piggyBac as a tool for genetic manipulation of human parasites, including Plasmodium falciparum. Data and materials availability: All data and code to understand and assess the conclusions of this research are available in the main text and supplementary materials and are deposited into the European Nucleotide Archive. Accession numbers for individual experiments and libraries are listed in table S9. piggyBac transfection plasmids (MRAS11/S2) and mutants are deposited with the Malaria Research Reagent and Reference Repository (BEI Resources), piggyBac transfection plasmids and mutant parasites may be obtained directly from the authors at the University of South Florida, using a standard academic materials transfer agreement based on the Uniform Biological Material Transfer Agreement.

SUPPLEMENTARY MATERIALS
www.sciencemag.org/content/360/6388/eaap7487/suppl/DC1

Corrected 3 May 2018. See full text.
Uncovering the essential genes of the human malaria parasite *Plasmodium falciparum* by saturation mutagenesis


*Science* 360 (6388), eaap7847.
DOI: 10.1126/science.aap7847

**Saturating malaria mutagenesis**

Malaria is caused by eukaryotic *Plasmodium* spp. parasites that classically infect red blood cells. These are difficult organisms to investigate genetically because of their AT-rich genomes. Zhang *et al.* have exploited this peculiarity by using piggyBac transposon insertion sites to achieve saturation-level mutagenesis for identifying and ranking essential genes and drug targets (see the Perspective by White and Rathod). Genes that are current candidates for drug targets were identified as essential, in contrast to many vaccine target genes. Notably, the proteasome degradation pathway was confirmed as a target for developing therapeutic interventions because of the several essential genes involved and the link to the mechanism of action of the current frontline drug, artemisinin.

*Science*, this issue p. eaap7847; see also p. 490