GeneTargeter: Automated in silico design for genome editing in the malaria parasite, *Plasmodium falciparum*

Pablo Cárdenas

Massachusetts Institute of Technology <https://orcid.org/0000-0001-7015-0512>

Lisl Y. Esherrick

Massachusetts Institute of Technology <https://orcid.org/0000-0002-4853-4007>

Gaël Chambonnier

Massachusetts Institute of Technology <https://orcid.org/0000-0002-5606-2089>

Sumanta Dey

Massachusetts Institute of Technology

Christopher V. Turlo

Massachusetts Institute of Technology

Armiyaw Sebastian Nasamu

Massachusetts Institute of Technology

Jacquin C Niles (✉ jcniles@mit.edu)

Massachusetts Institute of Technology <https://orcid.org/0000-0002-6250-8796>

Research Article

Keywords: malaria, functional genomics, computational design, genome editing, CRISPR/Cas9, homology-directed repair

Posted Date: June 21st, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-565539/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Version of Record: A version of this preprint was published at The CRISPR Journal on February 1st, 2022. See the published version at <https://doi.org/10.1089/crispr.2021.0069>.

Abstract

Functional characterization of the multitude of poorly-described proteins in the human malarial pathogen, *Plasmodium falciparum*, requires tools to enable genome-scale perturbation studies. Here we present GeneTargeter (genetargeter.mit.edu), a software tool for automating the design of homology-directed repair donor vectors to achieve gene knockouts, conditional knockdowns, and epitope tagging of *P. falciparum* genes. We demonstrate GeneTargeter facilitated genome-scale design of six different types of knockout and conditional knockdown constructs for the *P. falciparum* genome, and validate the computational design process experimentally with successful donor vector assembly. The software's modular, customizable nature makes it extendable to additional plasmids and genomes.

Introduction

Malaria, caused by parasites of the genus *Plasmodium*, puts about half the world's population at risk worldwide and is responsible for over 400,000 deaths every year [1]. Recent analyses anticipate a significant increase in malaria mortality for the year 2020 due to disruptions in control programs and antimalarial drug supply caused by the ongoing COVID-19 pandemic [2]. The spread of resistance to the already sparse number of approved antimalarial drugs [3–7] makes discovery of new drug candidates an urgent priority, but this effort has been hindered by a relatively poor understanding of *Plasmodium* biology [8–10]. Substantial improvements in the toolkit for studying gene function in *Plasmodium* have been made recently, including near genome scale essentiality screens in the human parasite *Plasmodium falciparum* [11] and the rodent malaria model, *Plasmodium berghei* [10]. While these studies permit the important classification

of gene essentiality, they do not provide the nuanced conditional regulation required to enable detailed analysis of gene function. Some 3,946 parasite genes (7 genome) remain annotated as putative or unknown function [12]. This is acutely problematic when studying red blood cell infection stages responsible for the clinical manifestations of malaria, where deletion of essential parasite genes in a haploid genome results in non-viable parasites [10, 11].

To address this gap, a number of tools for conditionally regulating gene expression have emerged, including: dimerizable Cre recombinase (diCre) for conditional gene deletion [13] or protein mislocalization [14]; *glmS* ribozyme [15] and TetR-aptamer [16–18] systems for conditional post-transcriptional/translation control; and degron systems for inducible protein degradation [19]. These have all proven useful in detailing the functions of parasite genes involved in diverse facets of *Plasmodium* biology [13–19]. Implementation of these gene regulation tools in *P. falciparum* has been done in a mostly ad hoc manner, where individual researchers adapt local design principles to study their specific gene(s) of interest. Furthermore, knowledge of *P. berghei* genes, respectively, are required for normal blood stage parasite biology [10, 11] provides impetus for establishing standardized frameworks for scaling the application of conditional gene regulation technologies towards achieving detailed functional understanding of parasite biology at genome-scale.

The successful implementation of CRISPR/Cas-based genome engineering in *Plasmodium* [20, 21] opens both automated design and technical opportunities for achieving increasingly scalable application of conditional gene knockdown technologies. CRISPR/Cas systems use an endonuclease that is easily programmed with single guide RNAs (sgRNAs) to induce DNA double-strand breaks at specific target sites within the parasite's genome, which trigger efficient homology-directed repair [20, 21]. The simple base pairing rules governing sgRNA selection and homologous repair regions afford straightforward rules for automating the design of donor vectors intended to specifically modify gene loci to achieve either conditional knockdown or knockout of gene expression. We recently reported a facile DNA assembly framework that permits modular assembly of TetR aptamer-based knockdown and knockout donor vectors via CRISPR/Cas-based genome editing [18]. The requirement for assembling long, AT-rich regions for homology-directed repair and transgene expression in *P. falciparum* make this an appealing DNA assembly framework amenable to automated construct design for increasingly scalable and high throughput functional genetics in *P. falciparum*.

Toward this goal, we have created GeneTargeter (genetargeter.mit.edu), a computational tool that fully automates donor vector design for conditional gene knockdown and knockout, while allowing users to adjust assembly parameters, when desired. The software outputs donor vector DNA sequence maps containing optimal designs for CRISPR/Cas modification of target genes, as well as oligo software provides custom DNA assembly instructions, explores multiple designs to identify the optimal choice, and provides feedback on predicted feasibility of a successful donor vector design to achieve editing. Overall, GeneTargeter's processing speed facilitates genome-scale design of *P. falciparum* donor vectors, permitting selection of designs for assembly on a scale suited to the intended functional studies.

Results

We created both a Web interface and a command line application (CLI) capable of running our computational design software. In both cases, the design algorithm for any given gene is the same, as summarized below.

Computational Design Process

GeneTargeter produces linear donor vector designs for gene knockout and knockdown using synthetic TetR aptamer-based post-transcriptional regulation [16–18]. Figure 1 summarizes the DNA elements (Fig. 1a–c) and regulatory mechanisms (Fig. 1e) GeneTargeter uses, and how these fit into the software's general operation and workflow (Fig. 1f,g).

For knockdown constructs, regulation is achieved via a TetR or TetR-DOZI protein fusion that reversibly interacts with a single or 10x TetR aptamer array positioned in the 5' UTR or 3' UTR, respectively, in an anhydrotetracycline (aTc)-dependent manner (Fig. 1 d,e). Genetically-encoded epitope tags can be introduced to facilitate immunodetection and immunoaffinity purification of target proteins. These constructs also include a multicistronic regulatory, reporter, and selection (RRS) module encoding TetR or TetR-DOZI (regulator), *Renilla* luciferase (reporter) and blasticidin S deaminase (selection marker) (Fig. 1e). For

knockout constructs, the TetR or TetR-DOZI RRS module can be used for positive selection. Target locus modification to achieve knockdown and knockout is achieved using CRISPR/Cas genome editing to induce site-specific, double-strand chromosomal DNA breaks to initiate homology-directed repair. The design of effective genome-wide knockdown and knockout vectors relies on scalable selection of both: (1) a suitable sgRNA for CRISPR/Cas induced cleavage; and (2) the appropriate DNA sequence upstream (left homologous region, LHR) and downstream (right homologous region, RHR) of the cleavage site to mediate homology-directed repair with concomitant insertion of DNA payloads for the desired outcomes (Fig. 1a–c).

For gene knockouts, the design process is relatively straightforward: selected sgRNAs are chosen from a region spanning the middle of the gene, and LHR and RHR selected to overlap generally with regions at or near the junction containing the start codon and 5' UTR (LHR) or stop codon and 3' UTR (RHR) of the gene. The design process for gene knockdowns is more involved. Ideally, the sgRNA will target within the 5' UTR or 3' UTR, as needed, and LHR and RHR sequences selected to flank the target cut site. Some genes only have sgRNA target sites within the coding sequence, however, which is addressed by inclusion of an additional recodonized region (RR) to prevent repeated locus cleavage after initial editing. The recodonized region is designed to: preserve the encoded protein's sequence; be immune to cleavage by multiple sgRNA selections; and allow for exclusion of intronic sequences, if desired. Given the highly AT-rich *P. falciparum* genome, the best predicted quality sgRNAs consistently target coding regions, thus designs satisfying the above criteria often include a recodonized region. Lastly, it is necessary to define the oligonucleotide and gene fragment sequences used to obtain the sgRNA, LHR and RHR for assembly into the appropriate base donor vector plasmid. Using as little as gene IDs as input, GeneTargeter automates selection of all sgRNA, LHR, RHR, primers for PCR and sequencing, and commercial gene synthesis fragments required for donor vector construction by Gibson Assembly [22]. The output also includes a post-editing target locus map to guide downstream analyses.

Application

GeneTargeter was used for *in silico* genome-scale design of knockdown (both 3'- and 5'-UTR modification) and knockout donor vectors targeting the *P. falciparum* 3D7 genome using Cas9 and Cas12a genome editing. Input gene files were obtained from the *P. falciparum* 3D7 reference genome available on the PlasmoDB database [12]. Successful designs were produced for between 4,118 – 4,942 (74–89%) of the 5,712 nuclear-encoded, protein-coding genes when using default settings (Fig. 2a).

A fraction of designs contained warnings, most commonly due to sub-optimal sgRNA and low primer complexity in non-coding regions. Cas12a consistently outperforms Cas9 in terms of number of successful designs, reflecting the high AT-richness of the *P. falciparum* genome and the Cas12a 5'-TTTV PAM. The web application processes genes at a rate of ≈ 1 gene/10 s, while total genome processing on a high-performance computing cluster took less than one hour. This is a significant improvement over designing each DNA construct and generating all output files manually, which can require 10–20 minutes of work per gene using the sgRNA and primer selection tools available on a DNA editing software suite such as Benchling (benchling.com). In addition, we present seven sample constructs designed by GeneTargeter using pSN054 vectors that have been assembled in the wet lab (Supplementary File S2). Successful assembly was confirmed in DNA gel electrophoreses of restriction enzyme digestion, PCR of individual components (Fig. 2b–h; Supplementary Tables S1, S2, and S3, and DNA sequencing.

Discussion

We implemented a modular and flexible computational framework for automated design of CRISPR/Cas-compatible homology-directed repair donor vectors towards facilitating genome-scale knockdown and knockout applications in the human malaria parasite, *P. falciparum*. Using a default set of user-adjustable parameters, donor vectors targeting most genes could be successfully designed (Fig. 2a). We demonstrate successful assembly of seven representative knockdown pSN054 constructs using GeneTargeter-predicted outputs [18] (Fig. 2b–h). In a minority of instances throughout the genome, no viable designs were predicted using default program settings (Fig. 2a). For Cas9-based constructs, this was most often due to lack of suitable sgRNA candidates based on an NGG PAM within the search region, resulting in design failures for 18 (0.33%), 16 (0.29%), and 5 (0.09%) genes, respectively, for 3' knockdown, 5' knockdown and knockout constructs. Cas12a-based designs were comparatively more successful given the higher frequency of Cas12a PAM sites in the genome, with design failures for 4 (0.07%), 6 (0.11%), and 4 (0.07%) genes, respectively, for 3' knockdown, 5' knockdown and knockout constructs. Given Cas12a's greater genomic coverage, the most common cause of failure in this case was selecting suitable homologous regions within closely adjacent transcriptional units. Failed designs occurred more commonly with noncoding RNAs or short peptides, which reduce the sgRNA search region and the probability of identifying a suitable sgRNA, and in regions of high gene density. Design failures resulting from limited sgRNA selection can be addressed by using a Cas enzyme with alternative PAM constraints [23]), which GeneTargeter allows. Along these lines, it is worth noting that 15 out of 18 of failed pSN054 designs based on Cas9 were covered by Cas12a designs, as were 14 out of 16 pSN150 knockdown designs and 2 out of 5 knockouts. Conversely, 1 out of 4 Cas12a pSN054 designs was covered by Cas9, as were 2 out of 6 pSN150 designs and 1 out of 4 knockouts. Finally, the software is more likely to produce design warnings in genes with repetitive sequences or exceptionally high (>95%) AT-content, such as with multi-intronic genes. The higher AT content of 5' compared with 3' UTRs also explains the slightly higher frequency of designs with warnings in 5' knockdown and knockout constructs when compared to 3' knockdowns, since this makes selecting recoded regions slightly more involved (Fig. 2a).

Conclusions

GeneTargeter provides researchers with a rapid design tool for a variety of genome editing needs in malaria parasites. Its modularity and flexibility with which design parameters can be adjusted facilitates expansion of the base donor vectors available for generating new donor vectors. With no additional adjustments to the software, homologous regions and sgRNAs designed for target genes can be used with arbitrary plasmids and DNA assembly strategies.

This is important as the TetR-aptamer system continues to be refined [18,24] and used more widely in *Plasmodium* research [24–46]. Furthermore, the same algorithms can be adjusted to accommodate *glmS* ribozyme regulation [15], degradation domains [19], diCre [14], and other approaches. Other generally used plasmid frameworks can be incorporated as well, such as those from the PlasmogEM resource [47, 48], or those used in the related Apicomplexan parasite, *Toxoplasma gondii* [49]. The GeneTargeter computational framework can easily be adapted for homology directed repair-based genome editing applications across a wide range of organisms, and can serve genome modification design needs of a broad research community.

Methods

Figure 3 summarizes how the different DNA elements needed for assembly are obtained from the original target gene sequence, as well as how primers and synthesized DNA fragments are designed to put together the final construct.

sgRNA design

sgRNA design criteria

The first step in the design algorithm is sgRNA selection. This is done by searching for instances of the chosen CRISPR enzyme's Protospacer Adjacent Motif (PAM) sequence, which defines possible sgRNAs. GeneTargeter supports PAM sequences for a variety of CRISPR enzymes, including *Streptococcus pyogenes* Cas9 (*SpCas9*) with a 3'-NGG PAM [50, 51] and *Acidaminococcus* sp. BV3L6 Cas12a (*AsCas12a*, previously known as Cpf1) with a 5'-TTTV PAM [52, 53].

Potential sgRNA sequences are evaluated based on their GC composition, presence of homopolymers, on-target scores and off-target scores. sgRNA GC richness is associated with better editing activity [54, 55]. Homopolymers have been reported as occasionally reducing sgRNA efficiency [54–56]. On-target scores, which gauge an sgRNA's predicted probability of cleaving the intended site, can be calculated with one of two algorithms: (1) the original Rule Set 2 score developed by Doench *et al.* for Cas9 sgRNAs [57] and updated as Azimuth (github.com/MicrosoftResearch/Azimuth); or (2) CINDEL scores, developed for use with Cas12a sgRNAs [53]. Finally, off-target activity scoring algorithms available include scores developed by Hsu *et al.* [58] and Cutting-Frequency Determination (CFD) scores developed by Doench *et al.* [57]. Both are based on aggregating the scores of pairwise comparisons between the sgRNA of choice and all other possible sgRNAs found in the target genome, with pairwise scores being determined using a coefficient matrix that takes mismatch position within the sgRNA into account, as well as mismatch identity in the case of CFD scores. Since the coefficient matrices for both scores were derived from *SpCas9* data in murine and human cells, an enzyme-agnostic score based on unweighted coefficients in the position mismatch matrix is provided for use as a proxy score with Cas12a sgRNAs. At the time of writing, we are aware of no off-target score specifically trained on data of Cas12a sgRNA activity.

sgRNA search sequence space

- The selection process begins by compiling a list of sgRNA candidates from a given sgRNA search space. The way this search space is defined depends on whether the desired construct is intended to modify the 3' or 5' UTR for knockdown, or achieve knockout of the target gene (Fig. 3a).
- *Knockdown via 3' UTR modification.* A sequence search space is defined by taking the last 450 bp of the target gene's coding sequence and 125 bp downstream of its stop codon. The limits of the search space can be modified by the user. The search space is then examined to identify possible PAMs. Preference is given to PAMs downstream of the stop codon, as this allows for straightforward omission of the sgRNA target site within homologous regions of the donor vector and no required modification of coding sequence. The software excludes sgRNA sequences found within genes downstream of the one being targeted.
- *Knockdown via 5' UTR modification.* A sequence search space is defined by taking 125 bp upstream of the target gene's start codon followed by the first 450 bp of coding sequence. The limits of the search space can be modified by the user. The search space is examined to identify possible PAMs. Preference is given to PAMs upstream of the start codon due to practicality, as explained above. The software excludes sgRNA sequences found within genes upstream of the one being targeted.
- *Knockout.* To maximize the probability of successful knockout and homologous recombination, potential sgRNAs in this case are selected from a search range centered around the midpoint of the coding sequence.

All potential sgRNA sequences within the search space are extracted based on their PAM to create a list of candidates ready for filtering.

sgRNA candidate filtering

Once the list of potential sgRNAs is defined, it is subjected to a primary filter step, wherein each criterion can be user-modified (Fig. 2a). Using default settings, a potential sgRNA is discarded if it has: 1) less than 25% GC content; 2) more than 10 consecutive A or T bases, which can noticeably reduce cleavage efficiency [55, 56]; or 3) an on-target score below a user-defined minimum threshold value, set to 35% and 20%, respectively, for Azimuth and CINDEL scores.

All sgRNAs passing the primary filtering step are subjected to a secondary series of tests, which include checking for any 4-homopolymers or triple T sequences [54, 55], and filtering according to off-target activity scores. Threshold off-target values empirically determined as useful are a minimum aggregated score of 20% and a maximum score of 50% for any pairwise hit in the case of CFD scores, and a 75% minimum aggregate score and 5% maximum pairwise hit score for Hsu *et al.* scores, all of which are user-adjustable. An sgRNA is discarded from the main candidate list if: 1) it contains 4-homopolymers or triple T sequences; 2) its aggregated score is below the set threshold; or 3) any pairwise score for a specific off-target sequence in the genome is above the set threshold. If an sgRNA fails this second series of tests, it is kept as a backup sgRNA.

Once three valid sgRNAs are found or the search area is exhausted, the sgRNA search process terminates. If no sgRNAs (valid or backup) were found, an error is written to the output message file and the vector design process aborts. If no valid sgRNAs are found, but there are backups, the backup sgRNA with the highest GC content is selected as the main sgRNA. In this case, a warning is written to the message file and the process continues. If there is at least one valid sgRNA, selection of the main sgRNA is as follows (Fig. 3a). If a valid sgRNA would not require a recoded region because it is outside the coding sequence, only valid and backup sgRNAs outside the coding sequence are kept. When there are no valid sgRNAs outside the coding sequence, the procedure to select sgRNAs depends on which end of the gene is being targeted.

- *Knockdown via 3' UTR modification.* The valid sgRNA closest to the stop codon is selected. All valid downstream sgRNAs and any up to 50 bp (user-adjustable) upstream of the selected sgRNA are kept. Backup sgRNAs that are downstream of the most upstream valid sgRNA are kept.
- *Knockdown via 5' UTR modification.* The valid sgRNA closest to the start codon is selected. All valid upstream sgRNAs and any up to 50 bp (user-adjustable) downstream of the selected sgRNA are kept. Backup sgRNAs that are upstream of the most downstream valid sgRNA are kept.
- *Knockout.* The valid sgRNA with highest GC content is selected. All valid and backup sgRNAs are kept.

The remaining valid sgRNAs are ranked and labeled (sgRNA 1, 2, 3) according to GC content, with the highest rank assigned based on highest GC content (Fig. 3a). All other sgRNAs are left unlabeled and not taken into account for the rest of the design process. Multiple sgRNAs are annotated and taken into account in order to facilitate simple exchange of the target sgRNA in a finished plasmid vector, in case the first choice proves to be experimentally unsuitable.

LHR design

GeneTargeter selects LHR and RHR regions to maximize the ease of PCR amplification and DNA assembly, while avoiding alterations to the gene coding sequence and its splicing patterns, as well as any adjacent coding and non-coding sequences.

LHR selection begins by selecting the 3' endpoint of the LHR. This occurs differently according to the type of vector being built (Fig. 3b):

- *Knockdown via 3' UTR modification.* The 3' end of the LHR is initially assigned to be immediately upstream of either the stop codon or the most upstream sgRNA target site, whichever is more upstream.
- *Knockdown via 5' UTR modification.* The 3' end of the LHR is assigned to either the start codon or the most upstream sgRNA, whichever is most downstream.
- *Knockout.* LHR selection follows that of 5' UTR-modifying knockdowns, providing an LHR close to the 5' end of the gene.

If the most upstream sgRNA target site is within an intron, the 3' end of the LHR is moved to the 3' end of the immediately upstream exon, and trimmed 3'→5' until the 3'-proximal 40 bp region of the LHR has a melting temperature $\leq 55^{\circ}\text{C}$ and is not in a repeat region (Fig. 3b). Melting temperature is calculated using nearest neighbor thermodynamics through BioPython [59]. If the 3' end of the LHR reaches ≤ 500 bp from the point where the search started and the melting temperature criterion is not satisfied, the LHR 3' end defaults to the initially assigned position, and a warning included with the returned design.

Next, the 5' end of the LHR is initially assigned to be 500 bp upstream of the selected 3' end. The 5' end of the LHR is moved upstream until the 5'-proximal 40 bp sequence of the LHR has a $\leq 55^{\circ}\text{C}$ and is not in a repeat region (Fig. 3b). If this melting temperature condition is not satisfied and the LHR exceeds 750 bp, the 5' end is instead moved downstream until the melting temperature criterion is satisfied or the LHR reaches 400 bp. If this also fails, the 3' end position will be moved upstream to the next suitable end point, and the 5' end position search is repeated from the beginning at this new location. This process iterates until both start and end positions meet melting temperature criteria or until the LHR search space is exhausted, as explained above. If a solution satisfying the melting temperature criterion is not found prior to exceeding the limits of the sequence space available, the initially assigned 5' end is selected and a warning is issued.

RHR design

GeneTargeter selects LHR and RHR regions to maximize the ease of PCR amplification and DNA assembly, while avoiding alterations to the gene coding sequence and its splicing patterns, as well as any adjacent coding and non-coding sequences.

- *Knockdown via 3' UTR modification.* The 5' end is defined to be immediately downstream of the target gene's stop codon or the most downstream sgRNA, whichever is more downstream.
- *Knockdown via 5' UTR modification.* The 5' end is defined to be immediately downstream of the most downstream sgRNA.
- *Knockout.* RHR selection follows that of 3' UTR-modifying knockdowns, providing an RHR close to the 3' end of the gene to delete as much of the gene coding sequence as possible.

If the most upstream sgRNA target site is within an intron, the 5' end of the RHR is moved to the 5' end of the immediately downstream exon, and trimmed 5'→3' until the 5'-proximal 40 bp region of the RHR has a melting temperature $\leq 55^{\circ}\text{C}$ and is not in a repeat region (Fig. 3c). If the 5' end of the RHR reaches ≤ 500 bp from the point where the search started and the melting temperature criterion is not satisfied, the RHR 5' end defaults to the initially assigned position, and a warning included with the returned design.

The 3' end of the RHR is defined and adjusted in the same way as the LHR 5' end. The 3' end of the RHR is initially assigned to be 500 bp downstream of the selected 5' end. The 3' end of the RHR is moved downstream until the 3'-proximal 40 bp sequence of the RHR has a melting temperature $\leq 55^{\circ}\text{C}$ and is not

in a repeat region (Fig. 3c). If this melting temperature condition is not satisfied and the RHR exceeds 750 bp, the 3' end is instead moved upstream until the melting temperature criterion is satisfied or the RHR reaches 400 bp. If this also fails, the 5' end position will be moved downstream to the next suitable end point, and the 3' end position search is repeated from the beginning at this new location. This process iterates until both start and end positions meet melting temperature criteria or until the RHR search is exhausted, as explained above. If a solution satisfying the melting temperature criterion is not found prior to exceeding the limits of the sequence space available, the initially assigned 3' end is selected and a warning is issued.

Recoded region design

Recoded regions are used to supplement missing coding sequences for certain knockdown constructs. They are defined differently according to vector type (Fig. 3d):

- *Knockdown via 3' UTR modification.* If the end of the LHR is within a coding region, GeneTargeter creates a recoded region spanning the end of the LHR to the end of the gene (excluding the stop codon).
- *Knockdown via 5' UTR modification.* If the start of the RHR is within a coding region, GeneTargeter creates a recoded region spanning the start of the gene (excluding the start codon) to the start of the RHR.
- *Knockout.* No recoded regions are necessary to knock out a gene.

If there are intronic regions within this sequence, the corresponding cDNA is used to synthesize a more compact recoded region. This is recodonized to preserve reading frame using *T. gondii* codon frequencies as the default to scramble sgRNA recognition sites and exclude restriction sites relevant during donor vector assembly. *T. gondii*, an apicomplexan parasite like *P. falciparum*, is selected because of its higher GC content, which facilitates both ease of synthesis and sequencing. Target sites of all ranked sgRNAs are recodonized to achieve pairwise CFD scores below 10%. This allows for modular exchange of sgRNAs in the final vector without the need to redesign and reassemble a new recoded region. Repetitive or homopolymeric sequences that may complicate assembly are recodonized. If the combined length of the recoded region and Gibson homology flanking sequences is below a default of 250 bp, the recoded region is extended until this minimum size requirement is met for compatibility with standard commercial synthesis.

Plasmid vector design

GeneTargeter virtually assembles a complete donor vector using a selected base plasmid with predetermined insertion sites for each component. Plasmid backbone pSN054 is used for 3' knockdowns, while plasmid pSN150 is used for both 5' knockdown and knockout constructs [18]. The software designs Gibson Assembly compatible primers for PCR amplification of the LHR, RHR and recoded regions. Designs are initialized with a preferred 40 bp homology on either side of the start and end positions of the region. If the initial primer has a melting temperature <50 °C, it is redesigned by increasing the length of the primer until the maximum primer length is achieved. The user is warned if the temperature difference between primer pairs is ≥5 °C. If needed, a recoded region gene fragment with 40 bp of homology to the plasmid backbone and either the LHR (for 3' UTR modification) or RHR (5' UTR modification) is designed, along with PCR primers for its amplification (Fig. 3e). All targeting elements and oligonucleotides are annotated both in the completed donor vector and edited gene locus output files.

Advanced methods and parameters

GeneTargeter can produce designs with minimal user intervention using only gene sequences input as GenBank files. However, the user may choose to manually specify one or multiple sgRNA sequences within the gene for the software to prioritize, or even the LHR or RHR sequences to use. In these cases, the software requires additional annotations for user-selected sgRNAs. The sgRNAs should contain "gRNA" in the annotation label and a number indicating prioritization order. Additionally, user-chosen annotations for the LHR and RHR (containing the capitalized words "LHR" and "RHR" in the labels, respectively) can be added to the file, in which case the software uses these custom regions instead of automatically selecting LHR and RHR. The user must specify the name of the annotation corresponding to the gene's coding sequence, if this name is different from the name of the file.

There are a variety of additional design parameters with default values that can be modified by the user (Table S4). At a global level, the user may select which plasmid to use, with options for gene knockout, 5' and 3' conditional knockdown. The user can also choose whether or not to include 5' epitope tags, and to do so only if the encoded protein does not have a predicted N-terminal trafficking signal. The user can also specify targets as "non-coding RNA" and exclusively search for editing sites flanking such regions.

Other adjustments can be made to homologous region selection, Gibson Assembly primer design and creation of recoded regions. The size of all homologous regions can be altered by manually adjusting minimum and maximum limits. Homologous region boundaries can be adjusted by changing the maximum allowable distance between the end of the LHR and start of the first sgRNA, and between the gene's stop codon and start of the RHR. The user may also define a range around the start and stop positions of each homologous region for the software to scan when searching for optimal positions, once a valid homologous region is identified. Furthermore, the sequence length at the start and end of each homologous region considered during melting temperature analysis and the melting temperature thresholds are both user-adjustable.

For Gibson Assembly primers, the user can define minimum, maximum and preferred lengths of homology between fragments. Each primer is twice this length, with half the primer binding to the amplified region and the other acting as an overhang designed to add the region of homology with the next fragment in the assembly. Other primer parameters that can be adjusted, including minimum and maximum melting temperature, and allowable differences in melting temperature between primer pairs.

Customization of sgRNA selection can be achieved by specifying the minimum GC content of gRNAs, the CRISPR cutting enzyme and PAM sequence being used, on-target and off-target scoring algorithms, and their respective threshold values. GeneTargeter allows for two different families of CRISPR enzymes,

Cas9 and Cas12, with variants covering fourteen different PAM sequences. On-target scoring algorithms include the Rule Set 2, Azimuth, and CINDEL scores, while off-target scores include Hsu (weighted and unweighted) and CFD algorithms.

The codon frequency table used for recodonization can be selected from a list including: apicomplexan *T. gondii* (default), *P. falciparum*, and *P. vivax*; non-apicomplexan model organisms *Escherichia coli*, *Saccharomyces cerevisiae*, and *Rattus norvegicus*; and a uniform-frequency, “codon scrambling” table. Two different codon optimization algorithms are available: Codon Adaptation Index (CAI) Maximization, in which each codon is replaced by its most common synonym according to the chosen frequency table [60], and Codon Sampling, in which codons are sampled randomly from their list of synonyms according to their individual relative frequency. In addition, the user can set threshold values for the minimum length of the recoded region.

Lastly, the user may specify the way GeneTargeter handles designs when the recoded region is under the minimum set threshold. The default course of action (i.e. extension of the recoded region) can be replaced by an alternative strategy in which oligonucleotides required for the preparation of the sgRNA sequence, along with 40 bp homology to be used in Gibson Assembly, through a primer annealing-extension reaction. If the length of the recoded region is less than 20 bp under this setting, the recoded region is included within the LHR reverse Gibson primer. The region will be designated for gene fragment synthesis regardless of the setting if its length exceeds the minimum gene fragment size.

Platforms and availability

GeneTargeter can run both as a web and command line application. In both cases, the software is composed of a number of separate, modular components handling different tasks (Fig. S1). At the core of the software is a Python script that processes input genes and produces output files. This core script draws on a series of both pre-existing and custom-built Python libraries, as well as external databases. These databases contain the empty vector constructs, sgRNA scoring databases [53, 57, 58], and codon frequency tables [61]. The user can interact with the software through a Web-based graphical user interface (GUI), available at genetargeter.mit.edu. This Web application was built using HTML5, CSS3, and JavaScript web technologies linked to a Python-run server. The server was built using Flask and Flask-SocketIO, and houses the main GeneTargeter scripts. Alternatively, the main GeneTargeter script can be run locally from a command line interface (CLI), useful if the user wishes to integrate the software within a larger bioinformatics pipeline or parallelize large batches of designs. The code for both the CLI and the web-based software can be found at github.com/pablocarderam/GeneTargeter. Input gene files for the CLI may be downloaded in bulk from PlasmoDB plasmodb.org (Supplementary File 3) for the case of the *P. falciparum* 3D7 genome.

Construct assembly and validation

DNA assemblies for pSN054-type constructs were carried out as explained in Nasamu *et al.* [18]. LHR and RHR DNA segments were obtained by PCR from genomic DNA isolated from the *P. falciparum* NF54 strain, while recoded regions and sgRNA cassettes were obtained by commercial synthesis. The individual fragments were introduced using Gibson Assembly [22] in two steps: first, for the RHR and the sgRNA cassette, and second for the LHR and recoded region. Assembly was confirmed by restriction enzyme digestions designed to reveal individual features of the plasmid, including the RHR and sgRNA cassette (AclI + I-SceI), LHR and Recoded Region (FseI + AsiSI), 10x TetR aptamer array alone (ApaI + XmaI), and 10x TetR aptamer array with PfHSP86 3' UTR (ApaI + XbaI). Digestion products were analyzed using 1% agarose gel electrophoresis. Construct assembly was additionally validated through Sanger DNA sequencing. All DNA primers and fragments required for Gibson assembly were designed by GeneTargeter (Supplementary Table S1).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The user can interact with the software through a Web-based graphical user interface (GUI), available at genetargeter.mit.edu. Alternatively, the software can be run locally from a command line interface (CLI). The code for both the CLI and the web-based software can be found at github.com/pablocarderam/GeneTargeter.

Competing interests

J.C.N. is listed as one of the inventors on a patent of the genetically encoded protein-binding RNA aptamer technology utilized. No other authors have competing interests to declare.

Funding

This work was supported by grants from the Bill and Melinda Gates Foundation (OPP1162467 and OPP1158199); Broad Next10; National Institute of General Medical Sciences Center for Integrative Synthetic Biology Grant (P50 GM098792); National Institutes of Environmental Health Sciences Core Center Grant (P30-ES002109) and National Institutes of Environmental Health Sciences Training Grant (T32-ES007020 (L.Y.E.)).

Authors' contributions

P.C., S.D., and J.C.N. conceptualized and designed the method. P.C. implemented the method. P.C., L.Y.E., G.C., S.D., and C.V.T. tested the method and iterated designs with J.C.N. A.S.N. contributed materials.

Acknowledgements

The authors would like to thank members of the Niles Lab for providing additional software testing, suggestions, and helpful feedback. The authors also thank Broad Institute Information and Technology Services for maintaining the Broad Univa GridEngine for Research (UGER) high performance computing cluster.

Author details

¹Department of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave 02139 Cambridge, MA, USA. ²Present address: Pfizer, Inc., Cambridge, MA, USA. ³Present address: Johns Hopkins Bloomberg School of Public Health, 615 Wolfe Street, Baltimore, MD 21205, USA. ⁴Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, Massachusetts 02142, USA

Additional Files

Additional file 1 — Supplementary figures and tables

A PDF file containing Supplementary Figure S1 and Supplementary Tables S1–4.

Additional file 2 — Sample assembled constructs

A zip file for a folder containing GeneTargeter files for 10 constructs implemented and transfected in the lab.

Additional file 3 — *P. falciparum* 3D7 genome input files

A zip file with all *P. falciparum* 3D7 genes in separate GenBank files formatted for use with GeneTargeter, identical to the one pre-loaded on the Web application. These files may be useful for users of the CLI application.

References

1. WHO: World malaria report 2020. World Health Organization (2019)
2. Organization, W.H., et al.: The potential impact of health service disruptions on the burden of malaria: a modelling analysis for countries in sub-saharan africa. Technical report, World Health Organization (2020)
3. Noedl, H., Se, Y., Schaecher, K., Smith, B.L., Socheat, D., Fukuda, M.M.: Evidence of artemisinin-resistant malaria in western cambodia. *New England Journal of Medicine* **359**(24), 2619–2620 (2008)
4. Ashley, E.A., Dhorda, M., Fairhurst, R.M., Amaratunga, C., Lim, P., Suon, S., Sreng, S., Anderson, J.M., Mao, S., Sam, B., et al.: Spread of artemisinin resistance in plasmodium falciparum malaria. *New England Journal of Medicine* **371**(5), 411–423 (2014)
5. Ye, R., Hu, D., Zhang, Y., Huang, Y., Sun, X., Wang, J., Chen, X., Zhou, H., Zhang, D., Mungthin, M., et al.: Distinctive origin of artemisinin-resistant plasmodium falciparum on the china-myanmar border. *Scientific reports* **6** (2016)
6. Saunders, D., Lon, C.: Combination therapies for malaria are failing—what next? *The Lancet Infectious Diseases* **16**(3), 274–275 (2016)
7. Imwong, M., Hien, T.T., Thuy-Nhien, N.T., Dondorp, A.M., White, N.J.: Spread of a single multidrug resistant malaria parasite lineage (pfprailin) to vietnam. *The Lancet Infectious Disease* **10**(23) (2017)
8. Florent, I., Maréchal, E., Gascuel, O., Brehelin, L.: Bioinformatic strategies to provide functional clues to the unknown genes in plasmodium falciparum genome. *Parasite* **283** (2010)
9. Webster, W.A., McFadden, G.I.: From the genome to the phenome: tools to understand the basic biology of plasmodium falciparum. *Journal of Eukaryotic Microbiology* **671** (2014)
10. Bushell, E., Gomes, A.R., Sanderson, T., Anar, B., Girling, G., Herd, C., Metcalf, T., Modrzynska, K., Schwach, F., Martin, R.E., et al.: Functional profiling of a plasmodium falciparum genome. *Nature* **546**, 272 (2017)
11. Zhang, M., Wang, C., Otto, T.D., Oberstaller, J., Liao, X., Adapa, S.R., Udenze, K., Bronner, I.F., Casandra, D., Mayho, M., et al.: Uncovering the essential genes of plasmodium falciparum. *Nature* **546**, 272 (2017)
12. Aurrecochea, C., Brestelli, J., Brunk, B.P., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., et al.: Plasmodb: a functional genomic database for plasmodium falciparum. *Nucleic Acids Research* **36**(13), 3827–3832 (2008)
13. Collins, C.R., Das, S., Wong, E.H., Andenmatten, N., Stallmach, R., Hackett, F., Herman, J., P., Müller, S., Meissner, M., Blackman, M.J.: Robust inducible cre recombinase activity in the human malaria parasite plasmodium falciparum enables efficient gene deletion within a single asexual erythrocytic growth cycle. *Molecular microbiology* **88**(4), 687–701 (2013)
14. Birnbaum, J., Flemming, S., Reichard, N., Soares, A.B., Mesén-Ramírez, P., Jonscher, E., Bergmann, B., Spielmann, T.: A genetic system to study plasmodium falciparum protein function. *Nature Methods* **14**(4), 450–456 (2017)

15. Prommana, P., Uthaipibull, C., Wongsombat, C., Kamchonwongpaisan, S., Yuthavong, Y., Knuepfer, E., Holder, A.A., Shaw, P.J.: Inducible knockdown of plas
16. Goldfless, S.J., Wagner, J.C., Niles, J.C.: Versatile control of plasmodium falciparum gene expression with an inducible protein-rna interaction. *Nature communications* **5**, 5329 (2014)
17. Ganesan, S.M., Falla, A., Goldfless, S.J., Nasamu, A.S., Niles, J.C.: Synthetic rna-protein modules integrated with native translation mechanisms to control gene expression in malaria parasites. *Nature communications* **7**, 10727 (2016)
18. Nasamu, A.S., Falla, A., Pasaje, C.F.A., Wall, B.A., Wagner, J.C., Ganesan, S.M., Goldfless, S.J., Niles, J.C.: An integrated platform for genome engineering and gene expression perturbation in plasmodium falciparum. *Scientific Reports* **11**(1), 1–15 (2021)
19. Armstrong, C.M., Goldberg, D.E.: An fbpk destabilization domain modulates protein levels in plasmodium falciparum. *Nature methods* **4**(12), 1007 (2007)
20. Wagner, J.C., Platt, R.J., Goldfless, S.J., Zhang, F., Niles, J.C.: Efficient crispr-cas9-mediated genome editing in plasmodium falciparum. *Nature methods* **11**(9), 915–918 (2014)
21. Ghorbal, M., Gorman, M., Macpherson, C.R., Martins, R.M., Scherf, A., Lopez-Rubio, J.-J.: Genome editing in the human malaria parasite plasmodium falciparum using the crispr-cas9 system. *Nature biotechnology* **32**(8), 819 (2014)
22. Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., Smith, H.O.: Enzymatic assembly of dna molecules up to several hundred kilobases. *Nature methods* **6**(5), 343–345 (2009)
23. Hu, J.H., Miller, S.M., Geurts, M.H., Tang, W., Chen, L., Sun, N., Zeina, C.M., Gao, X., Rees, H.A., Lin, Z., *et al.*: Evolved cas9 variants with broad pam compatibility and high dna specificity. *Nature* **556**(7699), 57–63 (2018)
24. Rajaram, K., Liu, H.B., Prigge, S.T.: A redesigned tet-aptamer system to control gene expression in plasmodium falciparum. *bioRxiv* (2020)
25. Sidik, S.M., Huet, D., Ganesan, S.M., Huynh, M.-H., Wang, T., Nasamu, A.S., Thiru, P., Saeij, J.P., Carruthers, V.B., Niles, J.C., *et al.*: A genome-wide crispr screen in toxoplasma identifies essential apicomplexan genes. *Cell* **166**(6), 1423–1435 (2016)
26. Nasamu, A.S., Glushakova, S., Russo, I., Vaupel, B., Oksman, A., Kim, A.S., Fremont, D.H., Tolia, N., Beck, J.R., Meyers, M.J., *et al.*: Plasmepsins ix and x are essential and druggable mediators of malaria parasite egress and invasion. *Science* **358**(6362), 518–522 (2017)
27. Spillman, N.J., Beck, J.R., Ganesan, S.M., Niles, J.C., Goldberg, D.E.: The chaperonin tric forms an oligomeric complex in the malaria parasite cytosol. *Cellul*
28. Amberg-Johnson, K., Hari, S.B., Ganesan, S.M., Lorenzi, H.A., Sauer, R.T., Niles, J.C., Yeh, E.: Small molecule inhibition of apicomplexan ftsh1 disrupts plastid biogen
29. Ke, H., Dass, S., Morrissey, J.M., Mather, M.W., Vaidya, A.B.: The mitochondrial ribosomal protein l13 is critical for the structural and functional integrity of th
- 8137 (2018)
30. Walczak, M., Ganesan, S.M., Niles, J.C., Yeh, E.: Atg8 is essential specifically for an autophagy-independent function in apicoplast biogenesis in blood-stage malaria parasites. *MBio* **9**(1) (2018)
31. Boucher, M.J., Ghosh, S., Zhang, L., Lal, A., Jang, S.W., Ju, A., Zhang, S., Wang, X., Ralph, S.A., Zou, J., *et al.*: Integrative proteomics and bioinformatic prediction enable a high-confidence apicoplast proteome in malaria parasites. *PLoS biology* **16**(9), 2005895 (2018)
32. Garten, M., Nasamu, A.S., Niles, J.C., Zimmerberg, J., Goldberg, D.E., Beck, J.R.: Exp2 is a nutrient-permeable channel in the vacuolar membrane of plasmodium and is essential for protein export via ptex. *Nature microbiology* **3**(10), 1090–1098 (2018)
33. Bhatnagar, S., Nicklas, S., Morrissey, J.M., Goldberg, D.E., Vaidya, A.B.: Diverse chemical compounds target plasmodium falciparum plasma membrane lipid homeostasis. *ACS infectious diseases* **5**(4), 550–558 (2019)
34. Istvan, E.S., Das, S., Bhatnagar, S., Beck, J.R., Owen, E., Llinas, M., Ganesan, S.M., Niles, J.C., Winzeler, E., Vaidya, A.B., *et al.*: Plasmodium niemann-pick type c1-related protein is a druggable target required for parasite membrane homeostasis. *Elife* **8**, 40529 (2019)
35. Rudlaff, R.M., Kraemer, S., Streva, V.A., Dvorin, J.D.: An essential contractile ring protein controls cell division in plasmodium falciparum. *Nature communica*: 13 (2019)
36. Tang, Y., Meister, T.R., Walczak, M., Pulkoski-Gross, M.J., Hari, S.B., Sauer, R.T., Amberg-Johnson, K., Yeh, E.: A mutagenesis screen for essential plastid biogenesis genes in human malaria parasites. *PLoS biology* **17**(2), 3000136 (2019)
37. Raj, D.K., Mohapatra, A.D., Jnawali, A., Zuromski, J., Jha, A., Cham-Kpu, G., Sherman, B., Rudlaff, R.M., Nixon, C.E., Hilton, N., *et al.*: Anti-pfgarp activates programmed cell death of parasites and reduces severe malaria. *Nature*, 1–5 (2020)
38. Ling, L., Mulaka, M., Munro, J., Dass, S., Mather, M.W., Riscoe, M.K., Llinas, M., Zhou, J., Ke, H.: Genetic ablation of the mitoribosome in the malaria parasite plasmodium falciparum sensitizes it to antimalarials that target mitor
- 7248 (2020)
39. Florentin, A., Stephens, D.R., Brooks, C.F., Baptista, R.P., Muralidharan, V.: Plastid biogenesis in malaria parasites requires the interactions and catalytic activ
40. Swift, R.P., Rajaram, K., Keutcha, C., Liu, H.B., Kwan, B., Dziedzic, A., Jedlicka, A.E., Prigge, S.T.: The ntp generating activity of pyruvate kinase ii is critical for apicoplast maintenance in plasmodium falciparum. *Elife* **9**, 50807 (2020)
41. Zimbres, F.M., Valenciano, A.L., Merino, E.F., Florentin, A., Holderman, N.R., He, G., Gawarecka, K., Skorupinska-Tudek, K., Fernandez-Murga, M.L., Swiezewska, E., *et al.*: Metabolomics profiling reveals new aspects of dolichol biosynthesis in plasmodium falciparum. *Scientific Reports* **10**(1), 020-70246-0
42. Nessel, T., Beck, J.M., Rayatpisheh, S., Jami-Alahmadi, Y., Wohlschlegel, J.A., Goldberg, D.E., Beck, J.R.: Exp1 is required for organisation of exp2 in the intraerythrocytic malaria parasite vacuole. *Cellular microbiology* **22**(5), 13168 (2020)
43. Fierro, M.A., Asady, B., Brooks, C.F., Cobb, D.W., Villegas, A., Moreno, S.N.J., Muralidharan, V.: An erccrec family protein regulates the egress proteolytic casca

44. Ramanathan, A.A., Morrissey, J.M., Daly, T.M., Bergman, L.W., Mather, M.W., Vaidya, A.B.: Oligomerization of the antimalarial drug target pfatp4 is essential f
45. Boucher, M.J., Yeh, E.: Evidence that disruption of apicoplast protein import in malaria parasites evades delayed-
death growth inhibition. *bioRxiv* (2018). doi:10.1101/422618. <https://www.biorxiv.org/content/early/2018/09/20/422618.full.pdf>
46. Cobb, D.W., Kudyba, H.M., Villegas, A., Hoopmann, M.R., Baptista, R., Bruton, B., Krakowiak, M., Moritz, R.L., Muralidharan, V.: A druggable oxidative folding |
47. Schwach, F., Bushell, E., Gomes, A.R., Anar, B., Girling, G., Herd, C., Rayner, J.C., Billker, O.: Plasmo gem, a database supporting a community resource for
large-scale experimental genetics in malaria parasites. *Nucleic acids research* **43**(D1), 1176–1182 (2015)
48. Pfander, C., Anar, B., Schwach, F., Otto, T.D., Brochet, M., Volkmann, K., Quail, M.A., Pain, A., Rosen, B., Skarnes, W., *et al.*: A scalable pipeline for highly effect
1082 (2011)
49. Sidik, S.M., Hackett, C.G., Tran, F., Westwood, N.J., Lourido, S.: Efficient genome engineering of toxoplasma gondii using crispr/cas9. *PloS one* **9**(6), 100450
50. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., Church, G.M.: Rna-
guided human genome engineering via cas9. *Science* **339**(6121), 823–826 (2013)
51. Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., Zhang, F.: Genome engineering using the crispr-cas9 system. *Nature protocols* **8**(11), 2281–
2308 (2013)
52. Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., *et al.*: Cpf1
guided endonuclease of a class 2 crispr-cas system. *Cell* **163**(3), 759–771 (2015)
53. Kim, H.K., Song, M., Lee, J., Menon, A.V., Jung, S., Kang, Y.-M., Choi, J.W., Woo, E., Koh, H.C., Nam, J.-W., *et al.*: In vivo high-throughput profiling of crispr-
cpf1 activity. *Nature methods* **14**(2), 153–159 (2017)
54. Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., *et al.*: Genome-
scale crispr-mediated control of gene repression and activation. *Cell* **159**(3), 647–661 (2014)
55. Hough, S.H., Kancleris, K., Brody, L., Humphries-
Kirilov, N., Wolanski, J., Dunaway, K., Ajetunmobi, A., Dillard, V.: Guide picker is a comprehensive design tool for visualizing and selecting guides for crispr e
56. Wang, T., Wei, J.J., Sabatini, D.M., Lander, E.S.: Genetic screens in human cells using the crispr-cas9 system. *Science* **343**(6166), 80–84 (2014)
57. Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., *et al.*: Optimized sgRNA design to r
target effects of crispr-cas9. *Nature biotechnology* **34**(2), 184–191 (2016)
58. Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., *et al.*: Dna targeting specificity of rna-guided
cas9 nucleases. *Nature biotechnology* **31**(9), 827–832 (2013)
59. Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., *et al.*: Biopython: freely available py
1423 (2009)
60. Puigb`o, P., Bravo, I.G., Garcia-Vallv`e, S.: E-
cai: a novel server to estimate an expected value of codon adaptation index (ecai). *BMC bioinformatics* **9**(1), 65 (2008)
61. Athey, J., Alexaki, A., Osipova, E., Rostovtsev, A., Santana-Quintero, L.V., Katneni, U., Simonyan, V., Kimchi-Sarfaty, C.: A new and updated resource for
codon usage tables. *BMC bioinformatics* **18**(1), 391 (2017)

Table

Table 1: GeneTargeter output file list

File name	File format	File contents
Pre-editing locus	GenBank (.gb)	A sequence file identical to the one supplied as input, containing the given gene annotated with the gRNAs, LHR, and RHR chosen automatically, if this option was selected.
Post-editing locus	GenBank (.gb)	A sequence file identical to the one supplied as input, containing the given gene after chromosomal editing has taken place. This file displays the gene and the fully-annotated synthetic regulatory payload inserted inside it within its genomic context.
Plasmid vector	GenBank (.gb)	A sequence file containing the fully-annotated plasmid vector designed to target the given gene. Annotations include oligonucleotides to be used for assembly and sequencing.
Oligo list	Comma-separated values (.csv)	A tabular file containing oligonucleotide sequences designed to construct the new plasmid.
Gene fragment list	FASTA file (.fasta)	A FASTA file containing gene fragment sequences designed to construct the new plasmid, ready to be ordered.
gRNA list	Comma-separated values (.csv)	A tabular file comparing all possible gRNAs evaluated by GeneTargeter before making a choice according to the scoring metrics and algorithm previously described, as well as the corresponding recoded gRNA for each in the final design.
Message log file	Text file (.txt)	A text file containing a message log and description of warnings or errors issued during the in silico design process.

Figures

Figure 1
GeneTargeter workflow Schematic of GeneTargeter's workflow. (a) An input file containing the annotated target gene(s) of interest (GOI) is selected or provided by the user with 1 kbp of DNA flanking its sequence. (b) GeneTargeter identifies the optimal guide RNA (sgRNA), left and right homology regions (LHR, RHR), and the appropriate codon-recoded region (RR) if necessary. (c) These sequences are incorporated into the parent vector plasmid, which is used to edit the native locus, resulting in (d) the post-editing locus for 3'-knockdown containing optional epitope tags, a regulatory aptamer array, selection marker, Renilla luciferase, the TetR-DOZI posttranslational regulator coding sequence, and sgRNA, depending on the knockdown or knockout context selected. (e) For conditional knockdown, the TetR or TetR-DOZI fusion protein binds to a single 5' aptamer or 3' aptamer array (illustrated), respectively, in the absence of anhydrotetracycline (aTc) to prevent translation. Adding aTc disrupts TetR binding to the aptamer and induces translation.

Figure 2

Method performance and example applications GeneTargeter performs with high success rates as evaluated both in the dry and wet lab. (a) Distribution of GeneTargeter automated design outputs for 5,516 protein-coding, nuclear genes in the *P. falciparum* 3D7 genome, classified according to whether they resulted in a successful design, a design with warning messages, or no valid designs using Cas9 and Cas12a as CRISPR cutting enzymes. The number of genes in each category is shown. (b–h) GeneTargeter generates predictions of the various sequences needed for creating a donor vector of interest. Illustrated here in the context of pSN054 V5 [18], is the physical assembly of 7 representative targets based on GeneTargeter outputs. For each gene target, the gel image on the left shows the PCR product for the amplified LHR and RHR. The gel image on the right shows the results of restriction digests performed to verify correct assembly into the pSN054 destination vector, namely: RHR + sgRNA cassette (Ascl + I-SceI); LHR + Recoded Region (FseI + AsiSI); 10x TetR aptamer array (Apal + Xmal); and 10x TetR aptamer array + PfHSP86 3' UTR (Apal + XbaI). (i) Map of a generic pSN054 construct shows the expected band sizes for each restriction digestion reaction. Designs for PF3D7_1304600 and PF3D7_1401800 contain additional XbaI cut sites within the LHR or recoded region, resulting in an additional band for the Apal + XbaI digestion. Primer sequences used and exact expected fragment sizes for restriction digests and PCRs are summarized in Tables S1–3.

Figure 3

Computational design process GeneTargeter automatically selects (a) the sgRNA, (b) left homologous region (LHR), (c) right homologous region (RHR), and (d) recoded region (if necessary) for a plasmid design targeting a given gene. (e) The software specifies the oligos and gene fragments needed to assemble donor vector plasmids based on their respective payloads shown in Fig. 1. The knockout design framework may also be used for “markerless” (i.e., no selection marker inserted during repair) genome editing, in which case the payload can be omitted.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [CardenasSupplementaryInformation20210606.docx](#)