MICROBIAL GENOMICS
Bases to Biology

Short Paper

EuPaGDT: a web tool tailored to design CRISPR guide RNAs for eukaryotic pathogens

Duo Peng[1,2,3] and Rick Tarleton[1,2,3]

[1]Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, GA 30602, USA

[2]Department of Cellular Biology, University of Georgia, Athens, GA 30602, USA

[3]Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

Correspondence: Rick Tarleton (tarleton@uga.edu)

DOI: 10.1099/mgen.0.000033

Recent development of CRISPR-Cas9 genome editing has enabled highly efficient and versatile manipulation of a variety of organisms and adaptation of the CRISPR-Cas9 system to eukaryotic pathogens has opened new avenues for studying these otherwise hard to manipulate organisms. Here we describe a webtool, Eukaryotic Pathogen gRNA Design Tool (EuPaGDT; available at http://grna.ctegd.uga.edu), which identifies guide RNA (gRNA) in input gene(s) to guide users in arriving at well-informed and appropriate gRNA design for many eukaryotic pathogens. Flexibility in gRNA design, accommodating unique eukaryotic pathogen (gene and genome) attributes and high-throughput gRNA design are the main features that distinguish EuPaGDT from other gRNA design tools. In addition to employing an array of known principles to score and rank gRNAs, EuPaGDT implements an effective on-target search algorithm to identify gRNA targeting multi-gene families, which are highly represented in these pathogens and play important roles in host–pathogen interactions. EuPaGDT also identifies and scores microhomology sequences flanking each gRNA targeted cut-site; these sites are often essential for the microhomology-mediated end joining process used for double-stranded break repair in these organisms. EuPaGDT also assists users in designing single-stranded oligonucleotides for homology directed repair. In batch processing mode, EuPaGDT is able to process genome-scale sequences, enabling preparation of gRNA libraries for large-scale screening projects.

Keywords: CRISPR-Cas9; eukaryotic pathogens; genome editing; gRNA design; webserver.

Abbreviations: CRISPR, clustered regularly interspaced short palindromic repeat; DSB, double-stranded break; gRNA, guide RNA; HDR, homology-directed repair; MMEJ, microhomology-mediated end joining; PAM, protospacer adjacent motif; ssODN, single-stranded oligonucleotide.

Data statement: We confirm that all supporting data, code and protocols have been provided within the article or through supplementary data files.

Introduction

RNA-guided Cas9 nuclease has enabled rapid, targeted modification of a wide range of genomes, including those of eukaryotic pathogens (Peng *et al.*, 2015; Shen *et al.*, 2014; Sidik *et al.*, 2014; Sollelis *et al.*, 2015; Zheng *et al.*, 2014), which are the causative agents of some of the most devastating and intractable diseases of humans. The CRISPR (clustered regularly interspaced short palindromic repeats)-Cas9 system is likely to be a particularly important tool for the study of gene function in pathogens that lack functional RNAi pathways, such as *Plasmodium* sp. (Ghorbal *et al.*, 2014; Lee & Fidock, 2014; Wagner *et al.*, 2014; Zhang *et al.*, 2014), the causative agent of malaria, and *Trypanosoma cruzi*, the agent of Chagas disease (Peng *et al.*, 2015). The CRISPR-Cas9 system has proven especially useful because of its relative ease of use and high efficiency as well as the ability to achieve multiple modifications per cell in a single organism. This latter property is particularly useful for modifying members of multigene families that are common in these pathogens (Peng *et al.*, 2015).

Received 10 July 2015; Accepted 14 September 2015

Impact Statement

Like other communities, the eukaryotic pathogen research community has begun to adopt the CRISPR-Cas9 system for engineering genomes of eukaryotic parasites at an unprecedented scale and ease. Our web-based Eukaryotic Pathogen gRNA Design Tool (EuPaGDT) enables researchers to arrive at well-informed guide RNA (gRNA) and homology-directed repair template design for CRISPR-Cas9 experiments in eukaryotic pathogens. Using an array of known principles, EuPaGDT characterizes all potential gRNAs in a user-input gene, allowing users to quickly identify refined gRNA in a gene sequence and greatly increases the chances of successful CRISPR-Cas9 experiments. EuPaGDT also features a batch processing mode, which can identify gRNA at a whole genome scale, allowing the *in silico* preparation of gRNA libraries for large-scale screening projects. EuPaGDT is currently available for 25 eukaryotic pathogen genomes, covering major lineages of eukaryotic pathogens and popular model organisms. Users can also upload custom genomes or request default genomes to be added to EuPaGDT. At this time, the EuPaGDT server has been running for 5 months, and has processed over 1340 requests from 654 users originating from 30 countries.

Current guide RNA (gRNA) design tools for the CRISPR-Cas9 system have limitations when applied to gRNA design for eukaryotic pathogens. For example, the genomes of most parasites exhibit great nucleotide sequence divergence (even at the within-species level) and harbour large, rapidly evolving gene families that are important players of host–pathogen interactions. To harness CRISPR-Cas9's multiplexing power to edit gene families, a gRNA design tool must handle multiple 'on-target' hits (gene family members) and discriminate them from true off-target sequences. To address these problems, we have developed a web tool, Eukaryotic Pathogen gRNA Design Tool (EuPaGDT), tailored to design gRNA for eukaryotic pathogens.

EuPaGDT

EuPaGDT identifies all possible gRNAs in an input sequence, and then calculates a ranked list of those gRNAs based on (1) on-target and off-target hit(s) in the selected or uploaded pathogen genome, (2) predicted gRNA activity and (3) identified microhomology pairs flanking the gRNA targeted cut site. Fig. 1 illustrates a workflow for a non-batch job request.

EuPaGDT can identify on-target hits in the genome for each gRNA, including those intended for editing multi-gene families. The on-target hit feature has the advantage of also identifying additional, sometimes unannotated copies of target genes. This feature is particularly important in incomplete and poorly annotated pathogen genomes, and/or in incompletely annotated gene family sets that can number in the thousands. To find all on-target genome-hits, the program compares the homology between sequences flanking the identified gRNA with all genome-target flanking sequences. If flanking homology regions meet the programmed identity criteria, then the genome-target will qualify as a potential on-target hit. Key parameters governing flanking homology comparison, such as sequence length and threshold for alignment identity and coverage, can be adjusted by the user to enable effective searching of on-targets in a spectrum of highly conserved to more divergent gene families. Identified similar sequences that do not meet these on-target criteria are automatically assigned to the off-target list. All genomic hits are annotated to aid the user in determining whether off-targets are within coding regions and to verify genomic loci annotations for each on-target site (Fig. 2).

A target score is assigned to each identified gRNA. The target score reflects how well a gRNA can target on-target sites while avoiding off-targets and is calculated as follows:

$$target\ score = \frac{on\text{-}target\ index}{maximum\ on\text{-}target\ index} - \frac{off\text{-}target\ index}{on\text{-}target\ index}$$

$$in\ which: on\text{-}target\ index = number\ of\ perfectly\ matching\ on\ target\ hits + 0.5 \times (number\ of\ imperfectly\ matching\ on\text{-}target\ hits)$$

$$off\text{-}target\ index = number\ of\ perfectly\ matching\ off\text{-}target\ hits + 0.5 \times (number\ of\ imperfectly\ matching\ off\text{-}target\ hits)$$

The first term of the target score equation is a fraction that reflects how well the current gRNA hits on-targets compared with the theoretical maximum target index. The maximum on-target index represents the maximum on-target number that gRNAs in a given input sequence can have; for example, a single-locus two-allele gene will have 2 as the 'maximum on-target index', and a gene with five copies in the genome would have 5 as the 'maximum on-target index'. The first term of the equation has a maximum value of 1. The second term of the formula evaluates how well the current gRNA avoids hitting off-target sites. Briefly, if a gRNA has far more on-targets than off-targets, the second term would be a small fraction, which is deducted from the first term; in contrast, the second term will be >1 if a gRNA has more off-targets than on-targets, giving a negative target score. The target score equation as a whole works equally well for single-locus genes and large gene families.

To further assess the potential utility of each gRNA identified, we have adopted a gRNA activity prediction scoring matrix empirically determined by Doench *et al.* (2014). The scoring matrix scores gRNA based on the nucleotide

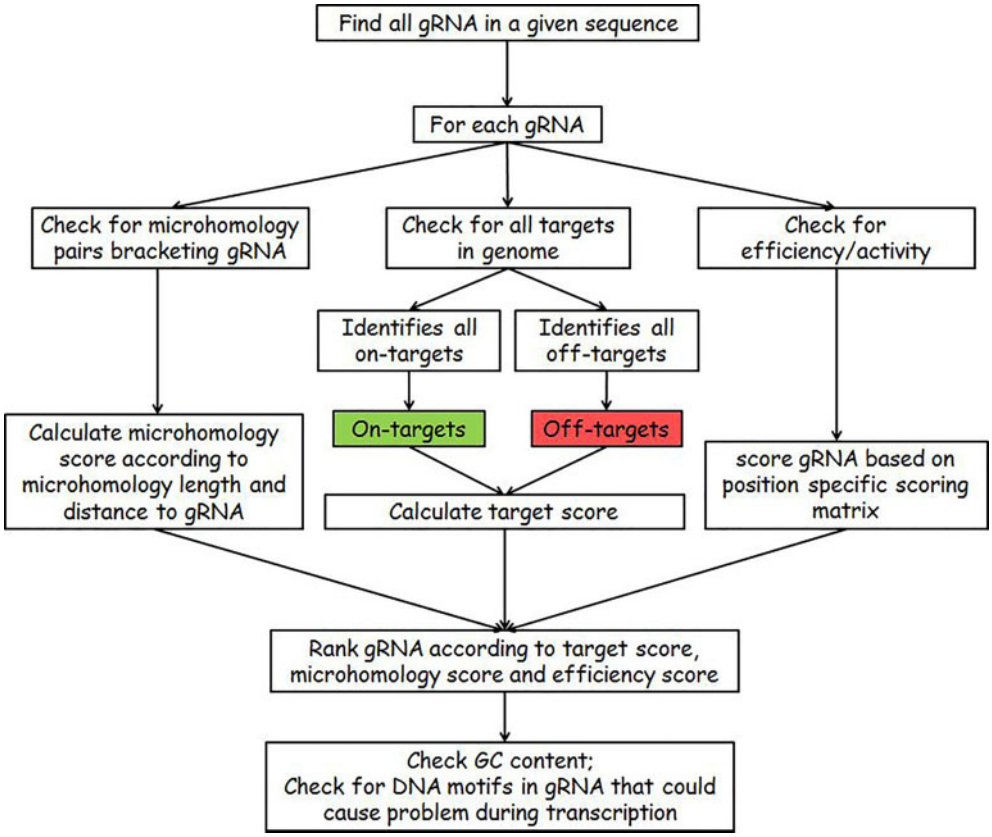


Fig. 1. Example workflow of a non-batch job request.

composition of gRNA's 20 nt targeting sequence as well as 4 and 3 nt up- and downstream, respectively. Although the gRNA activity prediction scoring matrix is developed from data obtained in mammalian cells, we have found in *T. cruzi* that a higher activity score (>0.5) correlates strongly with successful gRNA function ($n>10$).

EuPaGDT is extremely flexible, allowing users to tune many parameters governing the design of gRNA and characterization of on-/off-targets. Recent engineering of CRISPR-Cas9 nucleases has identified several smaller Cas9 orthologues as well as ones with altered protospacer adjacent motif (PAM) preferences (Kleinstiver *et al.*, 2015). The expanding list of available PAMs greatly increases the number of potential gRNAs in a given nucleotide sequence, allowing researchers to edit genomes with higher precision. EuPaGDT by default is programmed to use 'NGG' as an on-target PAM and 'NAG' and 'NGA' as off-target PAMs for the most widely used Cas9 from *Streptococcus pyrogenes* (SpCas9). However, users can specify multiple, custom on-target PAMs. For example, using the standard International Union of Pure and Applied Chemistry code (Cornish-Bowden, 1985) to specify degenerate PAM(s) of 3–10 bp, users can input a degenerative PAM sequence 'NNGRRT' recognized by Cas9 from *Staphylococcus aureus* (SaCas9) plus a variant PAM sequence 'NGA' used by SpCas9, in addition to using the classical PAM 'NGG' for SpCas9. EuPaGDT also allows users to specify multiple off-target PAMs. For example, SpCas9 can recognize the 'NGA' PAM at a low level (Kleinstiver *et al.*, 2015); off-targets using such alternative PAMs can be evaluated individually and integrated into the selection process for gRNAs if desired.

Users can refer to the 'table of on-/off-targets' pages available from the main result output page for each job request to visually inspect each on-target/off-target and their corresponding PAMs, as well as the alignment of gRNA with the target sequence. EuPaGDT provides other customized parameters such as gRNA length [as shorter gRNAs are shown to have fewer off-targets (Fu *et al.*, 2014)] and on-target and off-target search parameters can be custom specified to accommodate users' unique needs. For example: (1) searching for on-targets in a fast evolving multigene family might require relaxation of on-target searching criteria, lowering the 'identity' and 'coverage' parameter values in small decrements to determine if gRNA on-target numbers increase with a steady number of off-targets; and (2) selecting a shorter gRNA length might require a corresponding decrease in the 'maximum

(a)

gRNA_name(GeneName_gRNAstartPosition)	gRNA_sequence	gRNA_match_start	gRNA_match_end	alignment_identity	genome annotation	match_chromosome	chromosome_match_start	chromosome_match_end
TcCLB.506799.10-FATP_265	GCTATGGCAGAGATCACACA|TGG							
		1	23	100.00	TcCLB.511907.110: fatty+acid+transporter+protein-like%2C+putative	TcChr26-P	204139	204161
		1	23	100.00	TcCLB.506799.10: fatty+acid+transporter+protein-like%2C+putative	TcChr26-S	204353	204375
TcCLB.506799.10-FATP_266	CTATGGCAGAGATCACACAT|GGG							
		1	23	100.00	TcCLB.511907.110: fatty+acid+transporter+protein-like%2C+putative	TcChr26-P	204140	204162
		1	23	100.00	TcCLB.506799.10: fatty+acid+transporter+protein-like%2C+putative	TcChr26-S	204354	204376
TcCLB.506799.10-FATP_294	CGTAAGCATGACACTTCTCA|AGG							
		1	23	100.00	TcCLB.511907.110: fatty+acid+transporter+protein-like%2C+putative	TcChr26-P	204168	204190
		1	23	100.00	TcCLB.506799.10: fatty+acid+transporter+protein-like%2C+putative	TcChr26-S	204382	204404
TcCLB.506799.10-FATP_302	TGACACTTCTCAAGGCATTT|AGG							
		1	23	100.00	TcCLB.511907.110: fatty+acid+transporter+protein-like%2C+putative	TcChr26-P	204176	204198
		1	23	100.00	TcCLB.506799.10: fatty+acid+transporter+protein-like%2C+putative	TcChr26-S	204390	204412
TcCLB.506799.10-FATP_305	CACTTCTCAAGGCATTTAGG|AGG							
		1	23	100.00	TcCLB.511907.110: fatty+acid+transporter+protein-like%2C+putative	TcChr26-P	204179	204201
		1	23	100.00	TcCLB.506799.10: fatty+acid+transporter+protein-like%2C+putative	TcChr26-S	204393	204415
TcCLB.506799.10-FATP_335	AATTTGTCGCTTGTGCGCAG|CGG							
		1	23	100.00	TcCLB.511907.110: fatty+acid+transporter+protein-like%2C+putative	TcChr26-P	204209	204231
		1	23	100.00	TcCLB.506799.10: fatty+acid+transporter+protein-like%2C+putative	TcChr26-S	204423	204445
TcCLB.506799.10-FATP_336	ATTTGTCGCTTGTGCGCAGC|GGG							
		1	23	100.00	TcCLB.511907.110: fatty+acid+transporter+protein-like%2C+putative	TcChr26-P	204210	204232
		1	23	100.00	TcCLB.506799.10: fatty+acid+transporter+protein-like%2C+putative	TcChr26-S	204424	204446
TcCLB.506799.10-FATP_345	TTGTGCGCAGCGGGAACATC|AGG							
		1	23	100.00	TcCLB.511907.110: fatty+acid+transporter+protein-like%2C+putative	TcChr26-P	204219	204241
		1	23	100.00	TcCLB.506799.10: fatty+acid+transporter+protein-like%2C+putative	TcChr26-S	204433	204455
TcCLB.506799.10-FATP_348	TGCGCAGCGGGAACATCAGG|TGG							
		1	23	100.00	TcCLB.511907.110: fatty+acid+transporter+protein-like%2C+putative	TcChr26-P	204222	204244
		1	23	100.00	TcCLB.506799.10: fatty+acid+transporter+protein-like%2C+putative	TcChr26-S	204436	204458
TcCLB.506799.10-FATP_360	ACATCAGGTGGTGTTTGAGA|TGG							
		1	23	100.00	TcCLB.511907.110: fatty+acid+transporter+protein-like%2C+putative	TcChr26-P	204234	204256
		1	23	100.00	TcCLB.506799.10: fatty+acid+transporter+protein-like%2C+putative	TcChr26-S	204448	204470

(b)

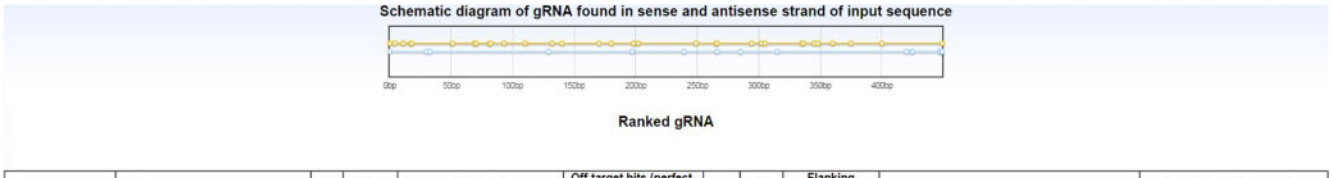
gRNA_name(GeneName_gRNAstartPosition)	gRNA sequence	gRNA match start	gRNA match end	alignment identity	genome annotation	match chromosome	chromosome match start	chromosome match end
TcCLB.506799.10-FATP_265	GCTATGGCAGAGATCACACA|TGG							
		9	23	100.00	TcCLB.508275.9: dynein+heavy+chain%2C+putative	TcChr30-P	T12682	T12696
TcCLB.506799.10-FATP_266	CTATGGCAGAGATCACACAT|GGG							
		no off-target hits						
TcCLB.506799.10-FATP_294	CGTAAGCATGACACTTCTCA|AGG							
		no off-target hits						
TcCLB.506799.10-FATP_302	TGACACTTCTCAAGGCATTT|AGG							
		7	23	94.12	TcCLB.508435.120: ATP-dependent+RNA+helicase%2C+putative	TcChr34-P	857218	857234
TcCLB.506799.10-FATP_305	CACTTCTCAAGGCATTTAGG|AGG							
		no off-target hits						
TcCLB.506799.10-FATP_335	AATTTGTCGCTTGTGCGCAG|CGG							
		no off-target hits						
TcCLB.506799.10-FATP_336	ATTTGTCGCTTGTGCGCAGC|GGG							
		no off-target hits						
TcCLB.506799.10-FATP_345	TTGTGCGCAGCGGGAACATC|AGG							
		7	23	94.12	TcCLB.511283.20: protein+transport+protein+Sec13%2C+putative	TcChr40-P	626927	626943
		7	23	94.12	TcCLB.506525.20: protein+transport+protein+Sec13%2C+putative	TcChr40-S	626266	626282
TcCLB.506799.10-FATP_348	TGCGCAGCGGGAACATCAGG|TGG							
		no off-target hits						
TcCLB.506799.10-FATP_360	ACATCAGGTGGTGTTTGAGA|TGG							
		no off-target hits						

Fig. 2. (a) Example output of genomic on-target hits and annotations for 10 gRNAs found in the TcFATP gene (gene id TcCLB.506799.10) in the *T. cruzi* CL Brener genome. (b) Example output of genomic off-target hits and annotations for 10 gRNAs found in the TcFATP gene in the *T. cruzi* CL Brener genome.

number of mismatches' allowed in the off-target parameter setting to ensure accurate off-target evaluation.

Microhomology-mediated end-joining (MMEJ) has been shown to repair double-stranded breaks (DSBs) generated by the CRISPR-Cas9 system in mammalian cells (Bae *et al.*, 2014; Wang *et al.*, 2013) and eukaryotic parasites (Peng *et al.*, 2015), often resulting in local sequence deletions which disrupt target genes. Analysing pairs of microhomology sequences that flank gRNA can help predict the size of MMEJ-mediated gene deletions. EuPaGDT identifies gRNA-flanking microhomology pairs of length 5–20 bp within 500 bp of each gRNA. Although MMEJ-mediated DSB repair is known to be involved in DSB repair in trypanosomes, the specific rules governing microhomology length and proximity to DSBs with respect to the efficiency of MMEJ DSB repair are not yet known (Glover *et al.*, 2011; Peng *et al.*, 2015). Therefore, EuPaGDT assigns each gRNA a score on a scale of 0–1 reflecting the length of microhomology pairs and their proximity to the gRNA-directed cut site, with a score of 1 for an ideal microhomology pair (>20 bp in length, and immediately flanking the gRNA cut site).

In addition to relying on MMEJ-induced deletions to mutate specific sequences, single-stranded oligonucleotides (ssODNs) bearing homology arms bracketing gRNA-guided DSBs can be used to introduce modifications at specific target positions by homology-directed repair (HDR; Wang *et al.*, 2013; Wu *et al.*, 2013; Yang *et al.*, 2013).



Ranked gRNA

gRNA id	gRNA sequence (PAM "NGG")	Total score	efficiency score	On-target hits (perfect-match | nonperfect-but-PAM-match)	Off-target hits (perfect-match | nonperfect-match)	Target score	GC content	Flanking microhomolgy pair score	Flanking microhomolgy pair score explanation	Potential problems during transcription
TcCLB.506799.10-FATP_110	TGCAGGCAAGAAACTCTATG|CGG	0.64	0.60	2 | 0	0 | 0	1.00	0.45	0.33	very good microhomology pair(s) found	gRNA does not start with G, please manually add a leading "G"
TcCLB.506799.10-FATP_375	TGAGATGGCGACAACAGATG|CGG	0.63	0.65	2 | 0	0 | 0	1.00	0.5	0.24	good microhomology pair(s) found	gRNA does not start with G, please manually add a leading "G"
TcCLB.506799.10-FATP_305	CACTTCTCAAGGCATTTAGG|AGG	0.63	0.60	2 | 0	0 | 0	1.00	0.45	0.29	good microhomology pair(s) found	gRNA contains >=3 "T(U)s" in a row, it will terminate POLIII transcription gRNA does not start with G, please manually add a leading "G"
TcCLB.506799.10-FATP_294	CGTAAGCATGACACTTCTCA|AGG	0.63	0.60	2 | 0	0 | 0	1.00	0.45	0.29	good microhomology pair(s) found	gRNA does not start with G, please manually add a leading "G"
TcCLB.506799.10-FATP_51	GCATTACGCATACATCACGA|CGG	0.62	0.60	2 | 0	0 | 0	1.00	0.45	0.25	good microhomology pair(s) found	No problem found
TcCLB.506799.10-FATP_335	AATTTGTCGCTTGTGCGCAG|CGG	0.61	0.65	2 | 0	0 | 0	1.00	0.5	0.19	good microhomology pair(s) found	gRNA contains >=3 "T(U)s" in a row, it will terminate POLIII transcription gRNA does not start with G, please manually add a leading "G"
TcCLB.506799.10-FATP_360	ACATCAGGTGGTGTTTGAGA|TGG	0.61	0.60	2 | 0	0 | 0	1.00	0.45	0.22	good microhomology pair(s) found	gRNA contains >=3 "T(U)s" in a row, it will terminate POLIII transcription gRNA does not start with G, please manually add a leading "G"
TcCLB.506799.10-FATP_170	TTGTGCCGTGCTGCAATTGC|TGG	0.60	0.61	2 | 0	0 | 0	1.00	0.55	0.20	good microhomology pair(s) found	gRNA does not start with G, please manually add a leading "G"
TcCLB.506799.10-FATP_336	ATTTGTCGCTTGTGCGCAGC|GGG	0.60	0.61	2 | 0	0 | 0	1.00	0.55	0.19	good microhomology pair(s) found	gRNA contains >=3 "T(U)s" in a row, it will terminate POLIII transcription gRNA does not start with G, please manually add a leading "G"
TcCLB.506799.10-FATP_266	CTATGGCAGAGATCACACAT|GGG	0.60	0.60	2 | 0	0 | 0	1.00	0.45	0.20	good microhomology pair(s) found	gRNA does not start with G, please manually add a leading "G"

Fig. 3. Example output of summary page of gRNA found in the TcFATP gene in the *T. cruzi* CL Brener genome (only the top 10 ranking gRNAs are shown).

Using such repair templates may also obviate undesired changes in the targeted gene (e.g. chromosome translocation or large-scale deletions; Brunet *et al.*, 2009; Cho *et al.*, 2014; Lee *et al.*, 2010). To aid researchers in designing oligonucleotide repair templates, EuPaGDT will by default generate an archetype ssODN sequence for each gRNA. Each ssODN repair donor has 30 bp of homology arms flanking gRNA's predicted DSB site and an 11 bp sequence consisting of three stop codons in three reading frames that will be inserted into the targeted sequence at the DSB site. Our default insertion site is at the DSB site (3 bp upstream of the PAM in SpCas9) because editing efficiency is highest when the desired nucleotide change is proximal to the DSB site (Bialk *et al.*, 2015; Yang *et al.*, 2013). Users can specify the length of the homology arms, and also enter the desired custom insertion-sequence in place of the default three-frame stop codon sequence. Further custom changes such as sequence deletion, nucleotide changes or insertion adjacent to the DSB site can be easily made at the user's discretion.

Additional quality control is performed by EuPaGDT to ensure that gRNA will be transcribed efficiently. EuPaGDT

gRNA id	gRNA sequence	oligo template for stop codon insertion
TcCLB.506799.10-FATP_110	TGCAGGCAAGAAACTCTATG|CGG	+strand:TCTTTGGGTGCCGTGCAGGCAAGAAACTCT**TAGATAGATAG**ATGCGGAACGCAAAGTTAATTTGTAAGGAT
		-strand:ATCCTTACAAATTAACTTTGCGTTCCGCAT**CTATCTATCTA**AGAGTTTCTTGCCTGCACGGCACCCAAAGA
TcCLB.506799.10-FATP_294	CGTAAGCATGACACTTCTCA|AGG	+strand:CACATGGGCAGTCCGTAAGCATGACACTTC**TAGATAGATAG**TCAAGGCATTTAGGAGGCAGTCACAATTTG
		-strand:CAAATTGTGACTGCCTCCTAAATGCCTTGA**CTATCTATCTA**GAAGTGTCATGCTTACGGACTGCCCATGTG
TcCLB.506799.10-FATP_305	CACTTCTCAAGGCATTTAGG|AGG	+strand:TCCGTAAGCATGACACTTCTCAAGGCATTT**TAGATAGATAG**AGGAGGCAGTCACAATTTGTCGCTTGTGCG
		-strand:CGCACAAGCGACAAATTGTGACTGCCTCCT**CTATCTATCTA**AAATGCCTTGAGAAGTGTCATGCTTACGGA
TcCLB.506799.10-FATP_375	TGAGATGGCGACAACAGATG|CGG	+strand:ATCAGGTGGTGTTTGAGATGGCGACAACAG**TAGATAGATAG**ATGCGGCCGTCCCAAAAATGAAGCTTTGTG
		-strand:CACAAAGCTTCATTTTTGGGACGGCCGCAT**CTATCTATCTA**CTGTTGTCGCCATCTCAAACACCACCTGAT
TcCLB.506799.10-FATP_51	GCATTACGCATACATCACGA|CGG	+strand:GGGTTATCATGTCGCATTACGCATACATCA**TAGATAGATAG**CGACGGCTGAGTATATAAAACTGGGATGCT
		-strand:AGCATCCCAGTTTTATATACTCAGCCGTCG**CTATCTATCTA**TGATGTATGCGTAATGCGACATGATAACCC
TcCLB.506799.10-FATP_335	AATTTGTCGCTTGTGCGCAG|CGG	+strand:AGGAGGCAGTCACAATTTGTCGCTTGTGCG**TAGATAGATAG**CAGCGGGAACATCAGGTGGTGTTTGAGATG
		-strand:CATCTCAAACACCACCTGATGTTCCCGCTG**CTATCTATCTA**CGCACAAGCGACAAATTGTGACTGCCTCCT
TcCLB.506799.10-FATP_360	ACATCAGGTGGTGTTTGAGA|TGG	+strand:GTGCGCAGCGGGAACATCAGGTGGTGTTTG**TAGATAGATAG**AGATGGCGACAACAGATGCGGCCGTCCCAA
		-strand:TTGGGACGGCCGCATCTGTTGTCGCCATCT**CTATCTATCTA**CAAACACCACCTGATGTTCCCGCTGCGCAC
TcCLB.506799.10-FATP_30_revcom	TCTTCCTACCGTATTGACGG|TGG	+strand:NNNNNNNNNNNNNNNNNATGGCGACCACCG**TAGATAGATAG**TCAATACGGTAGGAAGATGGAATTGGGTTA
		-strand:TAACCCAATTCCATCTTCCTACCGTATTGA**CTATCTATCTA**CGGTGGTCGCCATNNNNNNNNNNNNNNNNN
TcCLB.506799.10-FATP_170	TTGTGCCGTGCTGCAATTGC|TGG	+strand:GCATGGAAATTGATTGTGCCGTGCTGCAAT**TAGATAGATAG**TGCTGGAAGAAGAAGGAGAACTCTATTGTT
		-strand:AACAATAGAGTTCTCCTTCTTCTTCCAGCA**CTATCTATCTA**ATTGCAGCACGGCACAATCAATTTCCATGC
TcCLB.506799.10-FATP_266	CTATGGCAGAGATCACACAT|GGG	+strand:CATGATGTTGCCGCTATGGCAGAGATCACA**TAGATAGATAG**CATGGGCAGTCCGTAAGCATGACACTTCTC
		-strand:GAGAAGTGTCATGCTTACGGACTGCCCATG**CTATCTATCTA**TGTGATCTCTGCCATAGCGGCAACATCATG

Fig. 4. Example output of archetype ssODNs for the top 10 ranking gRNAs found in the TcFATP gene.

Flanking microhomologies for TcCLB.506799.10-FATP_69

color code:
best microhomolgy based on distance to gRNA
best microhomolgy based on length

Microhomolgy length	Microhomology sequence	Microhomology occurance position	gRNA position
5	ACGCA	56 134	69
5	CATCA	63 361	69
5	GCGAC	4 382	69
5	GGCGA	3 381	69
5	GCATT	51 316	69
5	TGTCG	47 339	69
5	ATCAC	64 277	69
5	TCATG	44 252	69
5	TGGGT	37 101	69
5	GGAAG	25 191	69
5	TGGCG	2 380	69
5	TAGGA	23 321	69
5	TTGGG	36 100	69
5	GTCGC	48 340	69
5	ATGGA	30 159	69
5	AGATG	28 377 390	69
5	GATGG	29 220 378	69
5	GAAGA	26 192 195	69
5	AATTG	34 164 184	69
5	ATGGC	1 268 379	69
5	ACATC	62 236 360	69
5	TGGAA	31 160 190	69
6	TGGCGA	2 380	69
6	TTGGGT	36 100	69
6	TGTCGC	47 339	69
6	GGAAGA	25 191	69
6	AGATGG	28 377	69
6	GGCGAC	3 381	69
6	ATGGAA	30 159	69
6	ATGGCG	1 379	69
6	ACATCA	62 360	69
7	ATGGCGA	1 379	69
7	TGGCGAC	2 380	69
8	ATGGCGAC	1 379	69

Fig. 5. Example output of microhomology pairs found for a gRNA in the TcFATP gene.

checks each identified gRNA for the presence of DNA motifs that may inhibit or terminate RNA polymerase transcription (Bogenhagen & Brown, 1981). EuPaGDT will also remind users to add a leading 'G' for efficient initiation of transcription when gRNA does not start with 'G'.

EuPaGDT ranks all gRNAs found in an input sequence based on their total score, which is calculated by unweighted averaging of the respective target score, activity-prediction score and microhomology-pair scoring. Our repeated test runs using a variety of input gene sequences show that the ranking process performs well in placing gRNAs with more desirable traits closer to the top of the list. Users can rapidly identify desirable gRNAs using the ranked list and further choose gRNA of interest based on specific usage, for example targeting a specific region of the input sequence.

For each request, EuPaGDT produces a summary page, including a schematic diagram showing each gRNA's position and strand in the input sequence, and a ranked list of such gRNAs along with concise results of each characterization step (Fig. 3). At a glance, users can easily grasp essential information relating to each gRNA. Additional detailed information [such as a summary of on-/off-targets (Fig. 2), archetype ssODNs (Fig. 4) and a summary of microhomology pairs (Fig. 5)] from each

characterization step is also available to the user to allow for selection of gRNA suitable to the project.

EuPaGDT also features a 'batch mode', which can process a list of genes and return a user-defined number of top-ranking gRNAs for each gene. Additionally, a multi-threaded batch mode is available to process genome-scale gene lists for genome-wide screening projects. EuPaGDT is currently available for 25 eukaryotic genomes, covering major lineages of eukaryotic pathogens as well as several popular model organisms. Users can also upload custom genomes or request default genomes to be available in EuPaGDT.

Informed and well-scrutinized gRNA design is instrumental to successful CRISPR-Cas9-mediated genome editing or gene expression manipulation. Here we describe a gRNA design web tool, EuPaGDT, tailored to eukaryotic pathogen gRNA design. By characterizing potential gRNAs with an array of currently accepted principles, EuPaGDT can facilitate researchers of eukaryotic pathogens to arrive at proper gRNA design for CRISPR-Cas9 experiments and thus bring new understanding for these otherwise difficult to manipulate pathogens.

Conclusion

We have developed EuPaGDT, a web-based tool, to assist the eukaryotic pathogen research community in designing gRNA and ssODN HDR templates for CRISPR-Cas9 experiments.

Acknowledgements

We thank Dave Dowless for technical assistance with setting up the webserver hosting EuPaGDT.

References

Bae, S., Kweon, J., Kim, H. S. & Kim, J. S. (2014). Microhomology-based choice of Cas9 nuclease target sites. *Nat Methods* **11**, 705–706.

Bialk, P., Rivera-Torres, N., Strouse, B. & Kmiec, E. B. (2015). Regulation of gene editing activity directed by single-stranded oligonucleotides and CRISPR/Cas9 systems. *PLoS One* **10**, e0129308.

Bogenhagen, D. F. & Brown, D. D. (1981). Nucleotide sequences in *Xenopus* 5S DNA required for transcription termination. *Cell* **24**, 261–270.

Brunet, E., Simsek, D., Tomishima, M., DeKelver, R., Choi, V. M., Gregory, P., Urnov, F., Weinstock, D. M. & Jasin, M. (2009). Chromosomal translocations induced at specified loci in human stem cells. *Proc Natl Acad Sci U S A* **106**, 10620–10625.

Cho, S. W., Kim, S., Kim, Y., Kweon, J., Kim, H. S., Bae, S. & Kim, J. S. (2014). Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res* **24**, 132–141.

Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res* **13**, 3021–3030.

Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B. L., Xavier, R. J. & Root, D. E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol* **32**, 1262–1267.

Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. (2014). Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol* **32**, 279–284.

Ghorbal, M., Gorman, M., Macpherson, C. R., Martins, R. M., Scherf, A. & Lopez-Rubio, J. J. (2014). Genome editing in the human malaria parasite *Plasmodium falciparum* using the CRISPR-Cas9 system. *Nat Biotechnol* **32**, 819–821.

Glover, L., Jun, J. & Horn, D. (2011). Microhomology-mediated deletion and gene conversion in African trypanosomes. *Nucleic Acids Res* **39**, 1372–1380.

Kleinstiver, B. P., Prew, M. S., Tsai, S. Q., Topkar, V. V., Nguyen, N. T., Zheng, Z., Gonzales, A. P., Li, Z., Peterson, R. T. & other authors (2015). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485.

Lee, M. C. & Fidock, D. A. (2014). CRISPR-mediated genome editing of *Plasmodium falciparum* malaria parasites. *Genome Med* **6**, 63.

Lee, H. J., Kim, E. & Kim, J. S. (2010). Targeted chromosomal deletions in human cells using zinc finger nucleases. *Genome Res* **20**, 81–89.

Peng, D., Kurup, S. P., Yao, P. Y., Minning, T. A. & Tarleton, R. L. (2015). CRISPR-Cas9-mediated single-gene and gene family disruption in *Trypanosoma cruzi*. *MBio* **6**, e02097–14.

Shen, B., Brown, K. M., Lee, T. D. & Sibley, L. D. (2014). Efficient gene disruption in diverse strains of Toxoplasma gondii using CRISPR/CAS9. *MBio* **5**, e01114–14.

Sidik, S. M., Hackett, C. G., Tran, F., Westwood, N. J. & Lourido, S. (2014). Efficient genome engineering of Toxoplasma gondii using CRISPR/Cas9. *PLoS One* **9**, e100450.

Sollelis, L., Ghorbal, M., MacPherson, C. R., Martins, R. M., Kuk, N., Crobu, L., Bastien, P., Scherf, A., Lopez-Rubio, J. J. & other authors (2015). First efficient CRISPR-Cas9-mediated genome editing in *Leishmania* parasites. *Cell Microbiol* **17**, 1405–1412.

Wagner, J. C., Platt, R. J., Goldfless, S. J., Zhang, F. & Niles, J. C. (2014). Efficient CRISPR-Cas9-mediated genome editing in *Plasmodium falciparum*. *Nat Methods* **11**, 915–918.

Wang, H., Yang, H., Shivalila, C. S., Dawlaty, M. M., Cheng, A. W., Zhang, F. & Jaenisch, R. (2013). One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910–918.

Wu, Y., Liang, D., Wang, Y., Bai, M., Tang, W., Bao, S., Yan, Z., Li, D. & Li, J. (2013). Correction of a genetic disease in mouse via use of CRISPR-Cas9. *Cell Stem Cell* **13**, 659–662.

Yang, L., Guell, M., Byrne, S., Yang, J. L., De Los Angeles, A., Mali, P., Aach, J., Kim-Kiselak, C., Briggs, A. W. & other authors (2013). Optimization of scarless human stem cell genome editing. *Nucleic Acids Res* **41**, 9049–9061.

Zhang, C., Xiao, B., Jiang, Y., Zhao, Y., Li, Z., Gao, H., Ling, Y., Wei, J., Li, S. & other authors (2014). Efficient editing of malaria parasite genome using the CRISPR/Cas9 system. *MBio* **5**, e01414–14.

Zheng, J., Jia, H. & Zheng, Y. (2014). Knockout of leucine aminopeptidase in *Toxoplasma gondii* using CRISPR/Cas9. *International journal for parasitology.*