CRISpy-Pop: A Web Tool for Designing CRISPR/Cas9-Driven Genetic Modifications in Diverse Populations

Hayley R. Stoneman,[*,†,‡,§] Russell L. Wrobel,[*,†,‡,§] Michael Place,[*,†] Michael Graham,[*,‡] David J. Krause,[*,†,‡,§] Matteo De Chiara,[**] Gianni Liti,[**] Joseph Schacherer,[††] Robert Landick,[*,§§,*] Audrey P. Gasch,[*,†,§] Trey K. Sato,[*,‡,§] and Chris Todd Hittinger[*,†,‡,§,1]**

[*]DOE Great Lakes Bioenergy Research Center, [†]Laboratory of Genetics, Center for Genomic Science Innovation, [‡]Wisconsin Energy Institute, [§]J. F. Crow Institute for the Study of Evolution, [§§]Department of Biochemistry,[***]Department of Bacteriology, University of Wisconsin-Madison, WI 53726, and [**]Université Côte d'Azur, CNRS, INSERM, IRCAN, Nice, France, [††]Université de Strasbourg, CNRS, GMGM UMR 7156, Strasbourg, France

ORCID IDs: 0000-0002-3796-2284 (H.R.S.); 0000-0002-5309-6718 (R.L.W.); 0000-0003-1014-350X (M.D.C.); 0000-0002-2318-0775 (G.L.); 0000-0002-5042-0383 (R.L.); 0000-0002-8182-257X (A.P.G.); 0000-0001-6592-9337 (T.K.S.); 0000-0001-5088-7461 (C.T.H.)

ABSTRACT CRISPR/Cas9 is a powerful tool for editing genomes, but design decisions are generally made with respect to a single reference genome. With population genomic data becoming available for an increasing number of model organisms, researchers are interested in manipulating multiple strains and lines. CRISpy-pop is a web application that generates and filters guide RNA sequences for CRISPR/Cas9 genome editing for diverse yeast and bacterial strains. The current implementation designs and predicts the activity of guide RNAs against more than 1000 *Saccharomyces cerevisiae* genomes, including 167 strains frequently used in bioenergy research. *Zymomonas mobilis*, an increasingly popular bacterial bioenergy research model, is also supported. CRISpy-pop is available as a web application (https://CRISpy-pop.glbrc.org/) with an intuitive graphical user interface. CRISpy-pop also cross-references the human genome to allow users to avoid the selection of guide RNAs with potential biosafety concerns. Additionally, CRISpy-pop predicts the strain coverage of each guide RNA within the supported strain sets, which aids in functional population genetic studies. Finally, we validate how CRISpy-pop can accurately predict the activity of guide RNAs across strains using population genomic data.

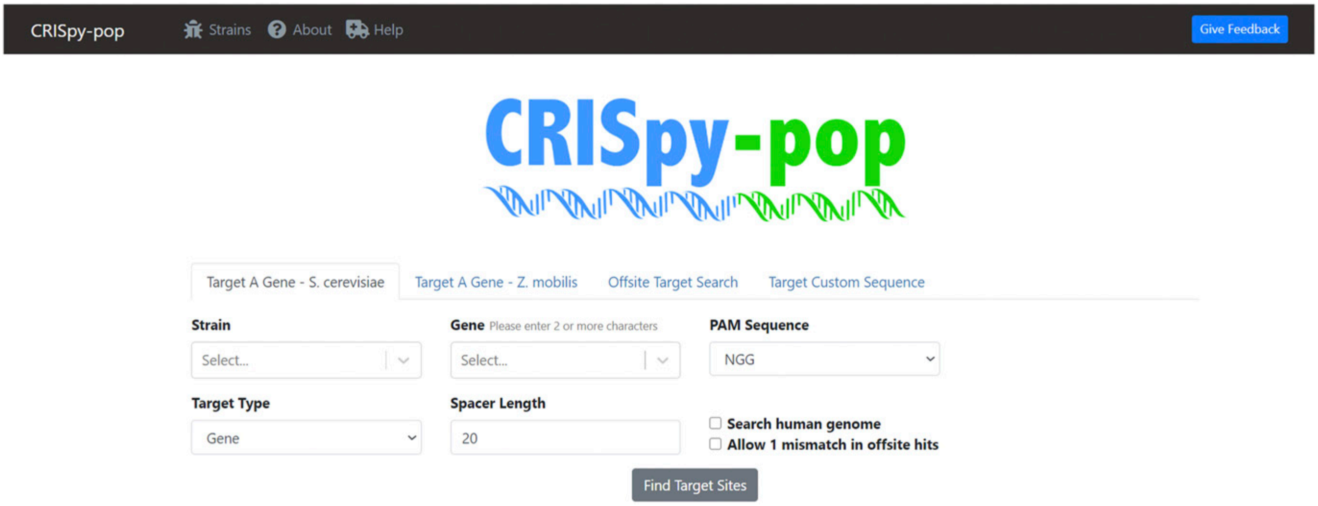
KEYWORDS
CRISPR/Cas9
sgRNA
genome editing
population genomics
yeast

CRISPR/Cas9 has become a widely used genome-editing tool due to its accuracy, precision, and flexibility (Ceasar *et al.* 2016). A primary step in designing a CRISPR/Cas9 experiment is the selection of the single guide RNA (sgRNA) target site, which usually occurs twenty nucleotides upstream of a protospacer adjacent motif (PAM) site. This sgRNA is bound by the Cas9 enzyme and used to direct Cas9 to the complementary location within the genome. Once bound to DNA, the Cas9 endonuclease cuts the DNA, leaving a double-strand break in the chromosome. This break can then be repaired using nonhomologous end-joining or homology-directed repair. Modified versions of Cas9 can nick a single DNA strand or bind to DNA sequence motifs without cleaving either DNA strand, while other Cas proteins have different sequence requirements (Makarova and Koonin 2015). For *Zymomonas mobilis*, there have been more published successes using Cas12a, which uses a different PAM site than Cas9 (Shen *et al.* 2019).

Due to the ease of manipulation of sgRNA targets by Cas9, it has been widely used for genome editing, including in yeasts and bacteria used in bioenergy research (Wang *et al.* 2016, Huang *et al.* 2016, Dong *et al.* 2016, Higgins *et al.* 2018, Kuang *et al.* 2018). The use of CRISPR/Cas9 for metabolic engineering in *S. cerevisiae* and *E. coli* has been extensively reviewed in Jakočiūnas *et al.* (2016). Besides its use in targeted gene knockouts (Stovicek, *et al.* 2015), it has also been used for targeted gene integrations (Ryan *et al.* 2014) and single nucleotide

doi: https://doi.org/10.1534/g3.120.401498
Manuscript received June 19, 2020; accepted for publication September 21, 2020; published Early Online September 21, 2020.

[1]Corresponding author: 1552 University Avenue, Madison, WI 53726. E-mail: cthittinger@wisc.edu



Strain: Select...

Gene Please enter 2 or more characters: Select...

PAM Sequence: NGG

Target Type: Gene

Spacer Length: 20

□ Search human genome

□ Allow 1 mismatch in offsite hits

Figure 1 Screenshot of the CRISpy-pop homepage (https://CRISpy-pop.glbrc.org/). There are options to search a gene in *S. cerevisiae* and *Z. mobilis*, as well as an offsite and custom target search. There are options to select specific strains, the desired PAM site, and the sgRNA length. Users may select the following PAM sites: NGG, NNGRRT, TTTV, NNNNGATT, TTTN, NCC, or NNAGAAW. Additionally, there is an option to search the human genome for perfect matches. CRISpy-pop features a user-friendly, web-based GUI.

changes (Higgins *et al.* 2018). With the use of nuclease-deficient Cas9 proteins fused to either transcriptional activation or interference domains, Lian *et al.* (2017) developed the CRISPR-AID system, which combines transcriptional activation, transcriptional interference, and gene deletion to combinatorically study perturbations in metabolic networks. By making it easier to design CRISPR/Cas9 tools, we aim to facilitate the engineering of yeast and *Zymomonas* strains to optimize the conversion of sugars from sustainably grown feedstocks into advanced biofuels and other bioproducts.

When designing sgRNAs, two main considerations must be made: efficiency and specificity. One tool that addresses the prediction of sgRNA efficiency is sgRNA Scorer 2.0 (Chari *et al.* 2017), which generated a model across multiple Cas9 orthologs to predict activity of sgRNAs from their sequence composition. A tool that addresses specificity of sgRNAs is Cas-OFFinder (Bae *et al.* 2014), which is a fast algorithm that searches specific genomes for potential off-target sites. Cui *et al.* 2018 reviewed a panel of twenty representative sgRNA design tools, which vary in their genome specificity, nuclease(s) supported, user input, and methods (or lack thereof) for on-target prediction and off-target scoring. Of the reviewed tools, eight supported the *Saccharomyces cerevisiae* reference genome and allowed for a variety of PAM sites. Of those eight, six provided both on-target prediction and off-target scoring, but none combined the use of sgRNA Scorer 2.0 and Cas-OFFinder into a single tool.

Recent advances in high-throughput sequencing have enabled the collection of population genomic data for an increasing number of organisms. Many studies have sequenced whole genomes of traditional and emerging model organisms, including large populations. For example, the 1000 Genomes Project (1000 Genomes Project Consortium *et al.* 2015) sequenced human genomes, the 1001 Genomes Project (1000 Genomes Project Consortium 2016) sequenced *Arabidopsis thaliana* genomes, and the 1002 Genomes Project (Peter *et al.* 2018) sequenced *S . cerevisiae* genomes. *S. cerevisiae* presents a high level of genetic diversity ($> 1\%$), more than 10 times greater than that found in humans. Interestingly, many of the detected genetic polymorphisms are low-frequency variants with almost 93% of the polymorphic sites associated with a minor allele frequency lower than 0.1 (Peter *et al.* 2018). Thus, the potential for strain-specific polymorphisms affecting sgRNA targeting is high.

With the increasing availability of population genomic data and the need to determine the functions of polymorphisms, there is a growing need to accommodate variation within species when designing CRISPR/Cas9-driven genetic modifications. Recently, the SNP-CRISPR tool was developed to address genomic variation by targeting single nucleotide polymorphisms (SNPs) (Chen *et al.* 2020). SNP-CRISPR supports several genetic model organisms, including humans, mouse, and *Drosophila melanogaster*, but it is limited to the variants included in user-supplied files, which creates a barrier for less computationally proficient users. The Yeastriction tool allows the user to choose between 33 different, commonly used yeast strains to find sgRNA target sites (Mans *et al.* 2015). However, this tool only searches for sgRNAs for the strain chosen and does not provide information about the variation of the sgRNA target sites between strains.

Here we developed and describe CRISpy-pop as a python-based (Van Rossum and Drake 2009) web application for the design of CRISPR/Cas9 sgRNAs for genetic modifications on populations of strains. CRISpy-pop incorporates popular diverse strain sets of *S. cerevisiae* from recent population genomic studies (Peter *et al.* 2018; Sardi *et al.* 2018) and uses the existing tools sgRNA Scorer 2.0 and Cas-OFFinder to assess the strain coverages of sgRNAs, predict their activities, and determine their off-target potentials. As a proof of principle, here we use CRISpy-pop to design *ade2* knockout mutants and accurately predict which strains can be targeted by which sgRNAs. CRISpy-pop fills a needed niche in functional and population genomic research.

MATERIALS AND METHODS

CRISpy-pop pipeline

The CRISpy-pop bioinformatic pipeline supports three modes of operation: targeting a gene, offsite target search, and targeting a custom sequence (Figure 1). We made use of open-source bioinformatic tools to generate sgRNA designs. The resulting sgRNA sequences are then scored and ranked based on predicted efficiency of the sgRNAs. Offsite target interactions are reported for each sgRNA

Table 1 Table of the oligonucleotides used. These include the bridge primers for adding the sgRNA sequences to the pKOPIS + sgRNA plasmid, the primers for PCR SOEing to clone the donor DNA, and the primers for PCR and Sanger sequencing

Name	Sequence
ADE2 Bridge L1	cgggtggcgaatgggactttACAGTTGGTATATTAGGAGGgttttagagctagaaatagc
ADE2 Bridge L2	cgggtggcgaatgggactttAACAGTTGGTATATTAGGAGgttttagagctagaaatagc
ADE2 Bridge H1	cgggtggcgaatgggactttACTTTGGCATACGATGGAAGgttttagagctagaaatagc
ADE2 Bridge H2	cgggtggcgaatgggactttACGGAGTCCGGAACTCTAGCgttttagagctagaaatagc
ADE2 5′ KO For	gatgtccacgacgtctctCAAATGACTCTTGTTGCATGG
ADE2 5′ KO Rev	GTATATCAATAAACTTATATAACTTGATTGTTTTGTCCGATTTTC
ADE2 3′ KO For	GAAAATCGGACAAAACAATCAAGTTATATAAGTTTATTGATATAC
ADE2 3′ KO Rev	cggtgtcggtgtcgtagGTATAATAAGTGATCTTATGTATG
ADE2 Conf For	ACCAACATAACACTGACATC
ADE2 Conf Rev	TATATGAACTGTATCGAAAC
pKOPIS sgRNA For	AACGCGAGCTGCGCACATAC
pKOPIS sgRNA Rev	GCGACAGTCACATCATGCC
pKOPIS sgRNA Seq For	CACCTATATCTGCGTGTTG
pKOPIS sgRNA Seq Rev	GCACGTCAAGACTGTCAAGG

across two complete strain sets of *S. cerevisiae*. These results are displayed to the user in a convenient and intuitive graphical user interface (GUI). The user can sort, search, save, and export the results in a more efficient way than would be possible using command line tools alone. For example, interactive sorting can be used to prioritize specific locations, high activities, or the numbers of strains targeted. CRISpy-pop also contains a genome viewer for visualization of each sgRNA within the target gene, facilitating design choices for the desired genome edits. CRISpy-pop supports a 167-strain set of *S. cerevisiae*, including 165 recently published genomes (Sardi *et al.* 2018), the S288C reference genome (Engel *et al.* 2014), and the GLBRCY22-3 bioenergy chassis (McIlwain *et al.* 2016); as well as a 1011-strain set of *S. cerevisiae* from the 1002 Yeast Genomes Project (Peter *et al.* 2018). CRISpy-pop's population genetic tool reports the numbers and identities of strains with perfect matches to each sgRNA. CRISpy-pop also contains a biosafety feature, which performs a local BLAST search (Altschul *et al.* 1990) of the human genome for perfect matches to each sgRNA sequence.

Targeting a gene in a specific strain: A mode was designed within CRISpy-pop to give the user the ability to target a gene by name in a specific strain. When the user selects this option, a streamlined search is performed to generate sgRNA sequences as follows. Gene coordinates are extracted from the appropriate GFF file. Using the reference genome FASTA, the gene sequence is extracted using samtools (Li *et al.* 2009). CRISpy-pop uses VCF files from its internal set of strains. These VCF files each contain variant calls for each strain relative to the S288C reference genome. These variants are then used to make substitutions in the S288C genome sequence to produce sequence files for each strain. This sequence is used as input to sgRNA Scorer 2.0 (Chari *et al.* 2017) in FASTA format with the appropriate PAM sequence and orientation (5′ or 3′) and the desired sequence length, which outputs a list of sgRNA sequences and their predicted activity scores. Cas-OFFinder (Bae *et al.* 2014) is then used to query all strains for offsite interactions, allowing zero mismatches; alternatively, the user may choose to allow one mismatch. The results are output in a user-friendly, graphical format. The results include a genome viewer, which shows the relative position of each sgRNA for the gene. In a table format, for each sgRNA, CRISpy-pop reports the sgRNA sequence, PAM site, activity score, GC%, chromosome, position, strand, position in the gene, mismatches, off-site matches, human genome hits, and strain coverage. Specific information can be obtained for any individual sgRNA by clicking on the desired entry in the table. For each sgRNA, an individual result report can be viewed, containing identities of strains predicted to be targeted, the alignment with the target, and the sgRNA details and statistics.

Offsite target search: For the offsite target search mode, a user provides a previously designed sgRNA sequence and selects the reference genome to be searched. Upon each search, CRISpy-pop employs Cas-OFFinder to provide a list of the specified reference genome's offsite targets for the user specified sgRNA. If no offsite targets exist, CRISpy-pop outputs that none were found.

Target a custom sequence: This mode allows the user to target a custom sequence, such as a gene that they may have previously engineered into a strain. When this feature is used, a custom DNA sequence is entered by the user. Once this sequence is entered, CRISpy-pop uses sgRNA Scorer 2.0 to find and score all potential sgRNAs within that sequence. Optionally, several supported reference genomes can be searched for offsite target matches, again using Cas-OFFinder. Currently supported genomes include *S. cerevisiae* S288C, *S. cerevisiae* GLBRCY22-3, *Saccharomyces paradoxus*, *Kluyveromyces lactis*, and *Zymomonas mobilis* ZM4. This tool outputs the same results as the gene target search.

Human hits search: CRISpy-pop performs a BLASTn database search of the human genome version hg38 (Schneider *et al.* 2017) for exact

Table 2 Table of the strains and plasmids used. These include the lab identifier used for each individual strain or plasmid. The strains include the reference, and the plasmids include the sgRNA target sequence

Strain	Lab Identifier	Reference
S288C	yHDO554	Mortimer and Johnston 1986
K1	yHEB306	Sardi *et al.* 2018
L1374	yHDPN448	Sardi *et al.* 2018
SK1	yHDPN454	Sardi *et al.* 2018
T73	yHDPN449	Sardi *et al.* 2018
Y55	yHDPN455	Sardi *et al.* 2018
Plasmid	**Lab Identifier**	**sgRNA Target Sequence**
L1	pHRW97	ACAGTTGGTATATTAGGAGG
L2	pHRW98	AACAGTTGGTATATTAGGAG
H1	pHRW104	ACTTTGGCATACGATGGAAG
H2	pHRW105	ACGGAGTCCGGAACTCTAGC

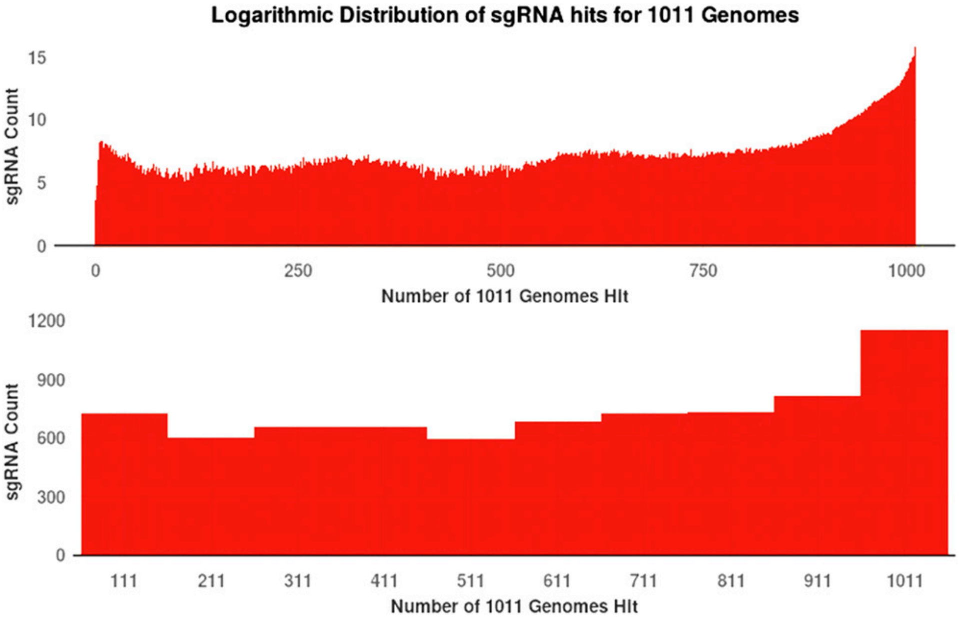


Figure 2 Log_2 histograms of sgRNAs found within 1011-strain set compared to S288C. The upper panel has a bin size of 1; the lower panel has a bin size of 100, except for the larger first bin. To explore sgRNAs designed against S288C using all non-mitochondrial verified ORFs *vs.* the variation within the 1011-strain set, we calculated the total number of strains in that set that could be targeted by each sgRNA found using S288C as the target. The total number of sgRNAs designed was 706,397. Only 55,875 of the sgRNAs had perfect matches in all 1011 genomes, while the remaining 605,522 target only the fraction of the genomes.

matches to each sgRNA as the query. If any perfect matches to the sgRNA are found, the output reports "Yes" under human hits.

Strain coverage function: This function uses the population genomic data described above to determine which strains are predicted to be targeted by each sgRNA. The strain coverage function searches the selected strain set for perfect matches to the sgRNA sequence and reports the number and identities of the strains covered.

Strain coverage function validation using ade2 mutants

ADE2 sgRNA selection: To validate CRISpy-pop's functionality, we used it to find sgRNAs to target the gene *ADE2* using a spacer length 20 and a PAM site of NGG. The 167-strain set (165 isolates, S288C and GLBRCY22-3) was searched for strain coverage. Two sgRNAs were selected with high strain coverage, and two were selected with low strain coverage, the latter of which were selected to have the exact same strain identities covered. The sgRNAs were chosen to balance the need for high activity scores, target more 5′ positions within the gene, and have no offsite matches.

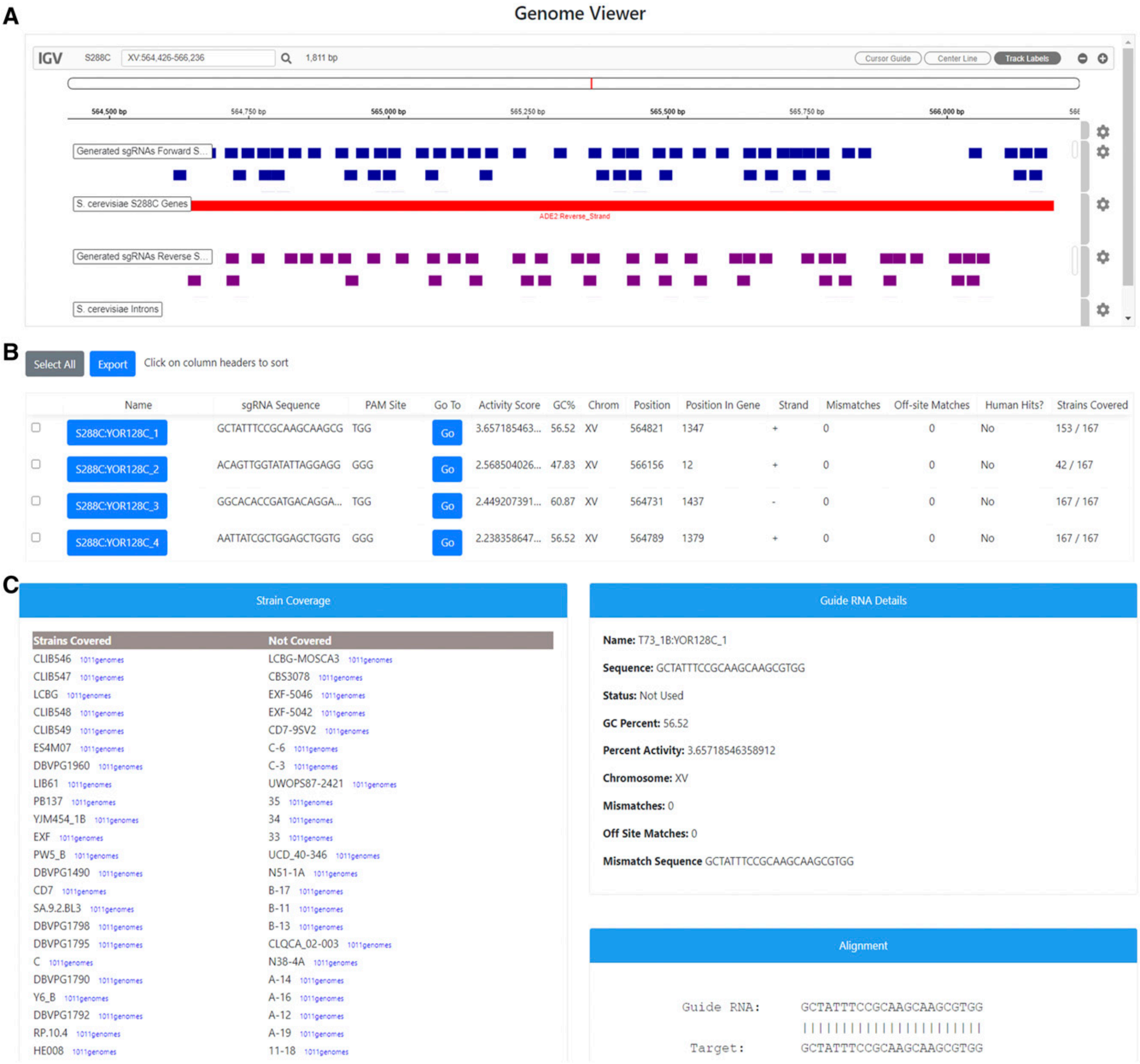
Plasmid and donor DNA synthesis: An empty sgRNA expression cassette, which contained the *SNR52* promoter, HDV ribozyme linked to a cloning site for sgRNA construct, and the *SNR52-1* terminator (Kuang *et al.* 2018), was first cloned into the pKOPIS plasmid (Kuang *et al.* 2018) using the NEBuilder HiFi DNA Assembly Master Mix (NEB #E2621) (Hsieh 2018). pKOPIS contains a *kanMX* selectable marker and encodes a Cas9 protein driven by the constitutive *RNR2* promoter. This empty pKOPIS + sgRNA plasmid (pHRW68) was linearized using a restriction enzyme digest with *Not*I.

Four different 60-nucleotide (nt), single-stranded bridging primers were designed, each containing one of the selected sgRNA sequences flanked by 20-nt homology regions with the pKOPIS plasmid (Table 1). The NEBuilder HiFi DNA Assembly Master Mix (Hseih 2018) was then used to clone the sgRNA sequences into the pKOPIS+ sgRNA plasmid, using the linearized plasmid and the bridge primers. This mixture was then used to transform *Escherichia coli* cells. The plasmids with the inserted sgRNAs were each isolated using the ZR Plasmid Miniprep Classic kit (Zymo Research) (Table 2). We confirmed correct sgRNA sequence insertion by performing BigDye (Applied Biosystems) Sanger-sequencing reactions with the pKOPIS sgRNA Seq primers.

The donor DNA was constructed using PCR splicing by overlap extension (SOEing) (Horton *et al.* 2013). All but the first 100 and last 100 base pairs of the gene were designed to be deleted from *ADE2*. A 40-nt primer was designed to amplify the 5′ forward portion of the gene and the homology region. A 40-nt primer was designed to amplify the 5′ reverse portion of the gene with 20-nt from the first 100 base pairs (bp) of the gene and 20-nt from the last 100 bp. The complement of this primer was then designed to amplify the 3′ forward portion of the gene. Finally, a 40-nt primer was designed for the 3′ reverse portion of the gene, which contained the last 20-nt of the gene and the homology region. Additionally, ADE2 Conf FOR and ADE2 Conf REV primers (Table 1) were used to confirm deletion of *ADE2* by PCR and sequencing.

The 3′ and 5′ sections of the donor DNA were first amplified individually using gradient PCR with annealing temperatures from 50° – 70° and Phusion High-Fidelity DNA Polymerase (New England Biolabs). The two individual sections were then joined into the complete donor DNA fragment using the same gradient PCR protocol. The final product was purified using the Qiagen QIAquick PCR Purification Kit according to the manufacturer's directions (Qiagen).

Transformation and knockout screening: All transformations were performed with pKOPIS plasmids containing each of the four sgRNAs or the empty vector as a negative control using the standard lithium acetate protocol optimized for *S. cerevisiae* (Gietz *et al.* 1995). In each reaction, 0.75 μg of sgRNA and 2 μg donor DNA were used. The transformations were grown in liquid YPD for three to five hours at 30° on a tissue culture rotator. They were then plated on three YPD + G418 (200μg/L) plates, with 100 μl, 200 μL, and 300 μL of transformation per plate. Successful transformants grew on YPD + G418 plates, while successful *ade2* knockouts also turned pink. The total number of colonies on each YPD + G418 plate was counted, as well as the number of pink colonies. The number of pink colonies was divided by the total number of successful transformants to calculate the efficiency of *ade2* deletion. Two pink colonies were chosen from



	Name	sgRNA Sequence	PAM Site	Go To	Activity Score	GC%	Chrom	Position	Position In Gene	Strand	Mismatches	Off-site Matches	Human Hits?	Strains Covered
□	S288C:YOR128C_1	GCTATTTCCGCAAGCAAGCG	TGG	Go	3.657185463...	56.52	XV	564821	1347	+	0	0	No	153 / 167
□	S288C:YOR128C_2	ACAGTTGGTATATTAGGAGG	GGG	Go	2.568504026...	47.83	XV	566156	12	+	0	0	No	42 / 167
□	S288C:YOR128C_3	GGCACACCGATGACAGGA...	TGG	Go	2.449207391...	60.87	XV	564731	1437	-	0	0	No	167 / 167
□	S288C:YOR128C_4	AATTATCGCTGGAGCTGGTG	GGG	Go	2.238358647...	56.52	XV	564789	1379	+	0	0	No	167 / 167

Strains Covered	Not Covered
CLIB546 1011genomes	LCBG-MOSCA3 1011genomes
CLIB547 1011genomes	CBS3078 1011genomes
LCBG 1011genomes	EXF-5046 1011genomes
CLIB548 1011genomes	EXF-5042 1011genomes
CLIB549 1011genomes	CD7-9SV2 1011genomes
ES4M07 1011genomes	C-6 1011genomes
DBVPG1960 1011genomes	C-3 1011genomes
LIB61 1011genomes	UWOPS87-2421 1011genomes
PB137 1011genomes	35 1011genomes
YJM454_1B 1011genomes	34 1011genomes
EXF 1011genomes	33 1011genomes
PW5_B 1011genomes	UCD_40-346 1011genomes
DBVPG1490 1011genomes	N51-1A 1011genomes
CD7 1011genomes	B-17 1011genomes
SA.9.2.BL3 1011genomes	B-11 1011genomes
DBVPG1798 1011genomes	B-13 1011genomes
DBVPG1795 1011genomes	CLQCA_02-003 1011genomes
C 1011genomes	N38-4A 1011genomes
DBVPG1790 1011genomes	A-14 1011genomes
Y6_B 1011genomes	A-16 1011genomes
DBVPG1792 1011genomes	A-12 1011genomes
RP.10.4 1011genomes	A-19 1011genomes
HE008 1011genomes	11-18 1011genomes

Name: T73_1B:YOR128C_1

Sequence: GCTATTTCCGCAAGCAAGCGTGG

Status: Not Used

GC Percent: 56.52

Percent Activity: 3.65718546358912

Chromosome: XV

Mismatches: 0

Off Site Matches: 0

Mismatch Sequence GCTATTTCCGCAAGCAAGCGTGG

```
Guide RNA:    GCTATTTCCGCAAGCAAGCGTGG
              |||||||||||||||||||||||
 Target:      GCTATTTCCGCAAGCAAGCGTGG
```

Figure 3 Sample output from CRISpy-pop searched for the gene *ADE2* in S288C genome with NGG PAM sequence, spacer length of 20, and cross-referencing the human genome to ensure no perfect matches exist for selected sgRNAs. A, genome viewer output by CRISpy-pop, showing the relative position of each sgRNA within the target gene. B, portion of the sgRNA table of results with each data point for each output sgRNA sequence. C, detailed results for an individual sgRNA, including identities of targeted and non-targeted strains.

each transformation of each strain, their *ADE2* genes were amplified by PCR, and their products were sequenced using Sanger sequencing to confirm that the knockouts had occurred using the donor DNA and homology-directed repair.

Statistical analysis

For the two strains targeted by all four sgRNAs (K1 and S288C), we used Mstat (https://mcardle.oncology.wisc.edu/mstat/) to calculate Kendall's Tau, performing a one-sided test for a correlation between the activity scores and the efficiencies.

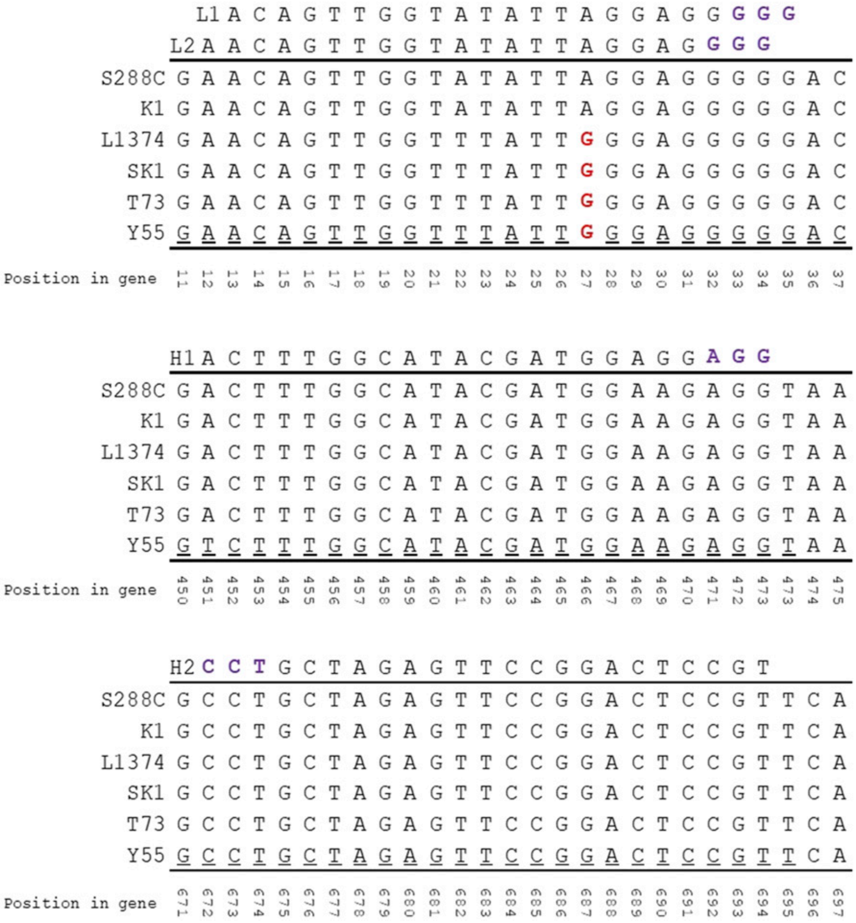
Data availability

CRISpy-pop is available online for non-commercial use at https://CRISpy-pop.glbrc.org/. The source code for the pipeline is available at: https://github.com/GLBRC/CRISpy-pop/ and https://github.com/GLBRC/crispy-pop-scripts. All new data generated is contained within this manuscript.

RESULTS AND DISCUSSION

Population-level variation in sgRNA target sites

S. cerevisiae is a useful genetic model system and bioengineering chassis due to its well-studied genome and ease of genetic manipulation. With growing population genomic datasets, functional investigations with CRISPR/Cas9 tools can now be extended beyond traditional laboratory strains, but variation in sgRNA target sites can still limit portability. To explore sgRNAs designed against S288C using all non-mitochondrial verified open reading frames (ORFs) *vs.* the variation found within the 1011-strain set, we calculated the total



```
          L1 A C A G T T G G T A T A T T A G G A G G G G G
        L2 A A C A G T T G G T A T A T T A G G A G G G G
     S288C G A A C A G T T G G T A T A T T A G G A G G G G G A C
        K1 G A A C A G T T G G T A T A T T A G G A G G G G G A C
     L1374 G A A C A G T T G G T T T A T T G G G A G G G G G A C
       SK1 G A A C A G T T G G T T T A T T G G G A G G G G G A C
       T73 G A A C A G T T G G T T T A T T G G G A G G G G G A C
       Y55 G A A C A G T T G G T T T A T T G G G A G G G G G A C
Position in gene 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37

        H1 A C T T T G G C A T A C G A T G G A G G A G G
     S288C G A C T T T G G C A T A C G A T G G A A G A G G T A A
        K1 G A C T T T G G C A T A C G A T G G A A G A G G T A A
     L1374 G A C T T T G G C A T A C G A T G G A A G A G G T A A
       SK1 G A C T T T G G C A T A C G A T G G A A G A G G T A A
       T73 G A C T T T G G C A T A C G A T G G A A G A G G T A A
       Y55 G T C T T T G G C A T A C G A T G G A A G A G G T A A
Position in gene 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 473 474 475

        H2 C C T G C T A G A G T T C C G G A C T C C G T
     S288C G C C T G C T A G A G T T C C G G A C T C C G T T C A
        K1 G C C T G C T A G A G T T C C G G A C T C C G T T C A
     L1374 G C C T G C T A G A G T T C C G G A C T C C G T T C A
       SK1 G C C T G C T A G A G T T C C G G A C T C C G T T C A
       T73 G C C T G C T A G A G T T C C G G A C T C C G T T C A
       Y55 G C C T G C T A G A G T T C C G G A C T C C G T T C A
Position in gene 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697
```

Figure 4 Portions of the *ADE2* gene from each strain aligned with the four sgRNAs. The PAM sites are included in purple. The *ADE2* gene sequence from each strain was extracted and aligned to each other and the four sgRNA sequences (H1, H2, L1, L2). The single nucleotide polymorphism highlighted in red at position 27 is predicted to prevent the two low-coverage sgRNAs (L1 and L2) from targeting *ADE2*. Note that sgRNA H2 targets the opposite strand, so its reverse complement is shown in this figure.

number of strains that could be targeted by each sgRNA using S288C as the design target. The total number of sgRNAs designed was 706,397. Only 55,875 of the sgRNAs had perfect matches in all 1011 genomes (Figure 2). Thus, randomly picking a sgRNA designed against the S288C reference genome would be unlikely to target all strains of potential interest. CRISpy-pop allows users to sort and filter by the number of strains targeted in a given gene, which aids design decisions to maximize sgRNA portability and facilitates population-level studies.

sgRNA selection using CRISpy-pop

To validate the strain coverage function of CRISpy-pop, we designed multiple sgRNAs targeting the gene *ADE2* with varying predicted strain coverage to create *ade2* knockout mutants in the Sardi *et al.* (2018) strain set (Figure 3). This strain set was chosen because we had access to the strains, but the 1011-strain population genomic data dataset (Peter *et al.* 2018) was also searched to compare relative strain coverage predictions for selected sgRNAs. Specifically, we selected two sgRNAs predicted to target all 167 strains (high-coverage sgRNAs) and two sgRNAs predicted to target only 42 of the 167 strains (low-coverage sgRNAs). The two high-coverage sgRNAs, H1 and H2, had activity scores of 1.341 and 0.426, respectively. The two low-coverage sgRNAs, L1 and L2, had activity scores of 2.569 and 2.050, respectively. None of the sgRNAs selected had any offsite matches or human hits. To determine whether the high-coverage sgRNAs also had high strain coverage within the previously published 1011-strain population genomic dataset, we reran the search with the same criteria on this dataset. H1 and H2 were also predicted to cut the vast majority of the 1011-strain set, targeting 905 and 910 genomes, respectively.

Yeast strain selection and transformations

We examined the strain coverage summary details from the CRISpy-pop search output for each sgRNA (Figure 3C) and selected six strains to test its predictive performance (Figure 4). Two strains (K1, S288C) were selected because they were predicted to be targeted by all four sgRNAs. These positive controls verified the functionality of all four sgRNAs and donor DNA constructs. The other four selected strains (L1374, SK1, T73, Y55) were predicted to be targeted by the high-coverage sgRNAs (H1 and H2) but not by the low-coverage sgRNAs (L1 and L2). S288C is haploid, while the other five strains are diploid.

Validation of sgRNA predictions made using CRISpy-pop

We transformed all six strains with CRISPR/Cas9 vectors expressing all four sgRNAs. We then counted the number of pink colonies, which are putative *ade2* knockouts due to deletion of *ADE2* causing the accumulation of aminoimidazole ribonucleotide (Silver and Eaton 1969), and we divided that number by the total number of transformants (G418-resistant) to calculate knockout efficiencies (Figure 5). The strains that were predicted to be targeted by all four sgRNAs were transformed first to ensure that all four sgRNAs were capable of producing *ade2* knockouts. All four sgRNAs successfully targeted the two predicted strains (K1 and S288C). We verified that homology-directed repair using the donor DNA - and not NHEJ - had

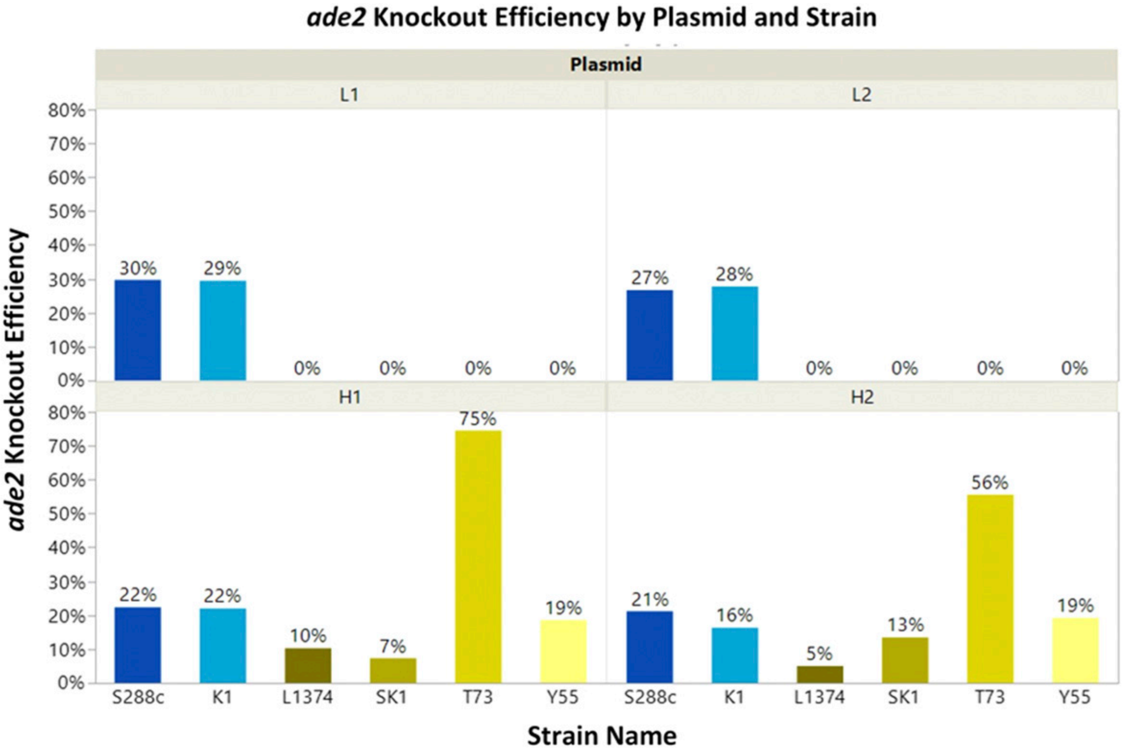


Figure 5 CRISpy-pop generated sgRNAs that target *ADE2* in a strain-specific manner. Results of the transformation of each strain with each sgRNA is shown. The two strains in red (S288C and SK1) were each predicted to be targeted by all four sgRNAs. Only these two strains both had non-zero % *ade2* knockouts (KOs) using all four sgRNAs. The four remaining strains were predicted to be targeted by only the high-coverage sgRNAs (H1 and H2), but not the low-coverage sgRNAs (L1 and L2). These four strains only had non-zero %*ade2* knockouts using the high-coverage sgRNAs. These results align with the strain coverage predictions made by CRISpy-pop. The predicted activity scores were H2 < H1 < L2 < L1, which are also consistent with the observed efficiencies.

occurred by Sanger-sequencing the *ADE2* PCR product. Once it was confirmed that all four sgRNAs could produce *ade2* knockouts using the donor DNA, the remaining strains were transformed with all four sgRNAs and donor DNA. As predicted, the low-coverage sgRNAs did not target the four strains (L1374, SK1, T73, and Y55) predicted to only be cut by the high-coverage sgRNAs, but the high-coverage sgRNAs all resulted in *ade2* knockout mutants.

Efficiencies varied widely by strain. For the two strains able to be cut by all four sgRNAs (K1 and S288C), the sgRNA activity scores predicted by CRISpy-pop correlated with their relative efficiencies (H2 < H1 < L2 < L1, Kendall's Tau = 0.85, P = 0.00101). These results validate the accuracy of the strain coverage and activity score predictions made by CRISpy-pop.

CONCLUSIONS

In summary, CRISpy-pop is a powerful and flexible design tool for planning and executing CRISPR/Cas9-driven genetic modifications on individual strains or large panels of strains. CRISpy-pop can continue be expanded to support new genomes as more data become available. The ability to target different PAM sites allows potential to use or screen for other Cas systems. It correctly predicts which strains can be targeted by which sgRNAs, as well as the activities of sgRNAs. Offsite targets, including a biosafety feature that scans for potential human genome binding, can be easily avoided with CRISpy-pop. This unique combination of features and its user-friendly web interface make CRISpy-pop ideal for designing experiments in diverse populations used for genetic engineering.

ACKNOWLEDGMENTS

We thank Kevin S. Myers for beta testing and feedback, Maria Sardi for providing datasets, Shihui Yang for suggesting the addition of some alternative PAM sites, and the University of Wisconsin Biotechnology Center DNA Sequencing Facility for providing Sanger-sequencing facilities and services. This material is based upon work supported in part by the Great Lakes Bioenergy Research Center, U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE-SC0018409; the National Science Foundation under Grant No. DEB-1442148; and the USDA National Institute of Food and Agriculture (Hatch Project 1020204). C.T.H. is a Pew Scholar in the Biomedical Sciences and a H. I. Romnes Faculty Fellow, supported by the Pew Charitable Trusts and Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation, respectively.

LITERATURE CITED

1001 Genomes Project Consortium 2016 1135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell 166: 481–491. .https://doi.org/10.1016/j.cell.2016.05.063

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

1000 Genomes Project Consortium; Auton, A., G. Abecasis, D. Altshuler *et al.*, 2015 A global reference for human genetic variation. Nature 526: 68–74. https://doi.org/10.1038/nature15393

Bae, S., J. Park, and J.-S. Kim, 2014 Cas-OFFinder: A fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics 30: 1473–1475. https://doi.org/10.1093/bioinformatics/btu048

Ceasar S.A., V. Rajan, S. Prykhozhij, J.N. Berman, S. Ignacimuthu, 2016 Insert, remove or replace: A highly advanced genome editing system using CRISPR/Cas9. Biochim Biphys Acta (BBA) 1863: 2333–2344. .https://doi.org/10.1016/j.bbamcr.2016.06.009

Chari, R., N. C. Yeo, A. Chavez, and G. M. Church, 2017 sgRNA Scorer 2.0 – a species independent model to predict CRISPR/Cas9 activity. ACS Synth. Biol. 6: 902–904. https://doi.org/10.1021/acssynbio.6b00343

Chen, C.-L., J. Rodiger, V. Chung, R. Viswanatha, S. E. Mohr *et al.*, 2020 SNP-CRISPR: A Web Tool for SNP-Specific Genome Editing. G3 (Bethesda) 10: 489–494. https://doi.org/10.1534/g3.119.400904

Cui, Y., J. Xu, M. Chen, X. Liao, and S. Peng, 2018 Review of CRISPR/Cas9 sgRNA Design Tools. Interdiscip. Sci. 10: 455–465. https://doi.org/10.1007/s12539-018-0298-z

Dong, G., M. He, and H. Feng, 2016 Functional Characterization of CRISPR-Cas System in the Ethanologenic Bacterium *Zymomonas mobilis* ZM4. Adv. Microbiol. 6: 178–189. https://doi.org/10.4236/aim.2016.63018

Engel, S. R., F. S. Dietrich, D. G. Fisk, G. Binkley, R. Balakrishnan *et al.*, 2014 The Reference Genome Sequence of *Saccharomyces cerevisiae:* Then and Now. G3 (Bethesda) 4: 389–398. https://doi.org/10.1534/g3.113.008995

Gietz, R. D., R. H. Schiestl, A. R. Willems, and R. A. Woods, 1995 Studies on the transformation of intact yeast cells by the LiAc/SS-DNA/PEG procedure. Yeast 11: 355–360. https://doi.org/10.1002/yea.320110408

Higgins, D. A., M. K. M. Young, M. Tremaine, M. Sardi, J. M. Fletcher *et al.*, 2018 Natural Variation in the Multidrug Efflux Pump SGE1 Underlies Ionic Liquid Tolerance in Yeast. Genetics 210: 219–234. https://doi.org/10.1534/genetics.118.301161

Horton, R. M., Z. Cai, S. N. Ho, and L. R. Pease, 2013 Gene Splicing by Overlap Extension: Tailer-Made Genes Using the Polymerase Chain Reaction. Biotechniques 54. https://doi.org/10.2144/000114017

Hsieh Peichung, 2018 Bridging dsDNA with a ssDNA Oligo and NEBuilder HiFi Assembly to create an sgRNA-Cas9 Expression Vector. New England Biolabs Inc. https://www.neb.com/-/media/nebus/files/application-notes/bridging_dsdna_with_ssdna_oligo_and_nebuilder_hifi_dna_assembly_to_create_sgrna-cas-9_expression_vector_an.pdf?la=en&rev=e8d96021187447e188161a0b48ee9800.

Huang, H., C. Chai, N. Li, P. Rowe, N. P. Minton *et al.*, 2016 CRISPR/Cas9-Based Efficient Genome Editing in *Clostridium ljungdahlii*, an Autotrophic Gas-Fermenting Bacterium. ACS Synth. Biol. 5: 1355–1361. https://doi.org/10.1021/acssynbio.6b00044

Jakočiūnas, T., M. K. Jensen, and J. D. Keasling, 2016 CRISPR/Cas9 advances engineering of microbial cell factories. Metab. Eng. 34: 44–59. https://doi.org/10.1016/j.ymben.2015.12.003

Kuang, M. C., J. Kominek, W. G. Alexander, J. F. Cheng, R. L. Wrobel *et al.*, 2018 Repeated Cis-Regulatory Tuning of a Metabolic Bottleneck Gene during Evolution. Mol. Biol. Evol. 35: 1968–1981. https://doi.org/10.1093/molbev/msy102

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lian, J., M. HamediRad, S. Hu, H. Zhao, 2017 Combinatorial metabolic engineering using an orthogonal trifunctional CRISPR system. Nat. Commun. 8: 1688. https://doi.org/10.1038/s41467-019-13621-4

Mans, R., H. M. van Rossum, M. Wijsman, A. Backx, N. G. A. Kuijpers *et al.*, 2015 CRISPR/Cas9: a molecular Swiss army knife for simultaneous introduction of multiple genetic modifications in Saccharomyces cerevisiae. FEMS Yeast Res. 15: fov004. https://doi.org/10.1093/femsyr/fov004

Makarova, K. S., and E. V. Koonin, 2015 Annotation and Classification of CRISPR-Cas Systems. Methods Mol. Biol. 1311: 47–75. https://doi.org/10.1007/978-1-4939-2687-9_4

McIlwain, S. J., D. Peris, M. Sardi, O. V. Moskvin, F. Zhan *et al.*, 2016 Genome Sequence and Analysis of a Stress-Tolerant, Wild-Derived Strain of *Saccharomyces cerevisiae* Used in Biofuels Research. G3 (Bethesda) 6: 1757–1766. https://doi.org/10.1534/g3.116.029389

Mortimer, R., and J. Johnston, 1986 Genealogy of principal strains of the yeast genetic stock center. Genetics 113: 35–43. https://www.genetics.org/content/113/1/35

Peter, J., M. De Chiara, A. Friedrich, J.-X. Yue, D. Pflieger *et al.*, 2018 Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. Nature 559: 339–344. https://doi.org/10.1038/s41586-018-0030-5

Ryan, O. W., J. M. Skerker, M. J. Maurer, X. Li, J. C. Tsai *et al.*, 2014 Selection of chromosomal DNA libraries using a multiplex CRISPR system. eLife. https://doi.org/10.7554/eLife.03703

Sardi, M., V. Paithane, M. Place, E. Robinson, J. Hose *et al.*, 2018 Genome-wide association across *Saccharomyces cerevisiae* strains reveals substantial variation in underlying gene requirements for toxin tolerance. PLoS Genet. 14: e1007217. https://doi.org/10.1371/journal.pgen.1007217

Schneider, V. A., T. Graves-Lindsay, K. Howe, N. Bouk, H.-C. Chen *et al.*, 2017 Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27: 849–864. https://doi.org/10.1101/gr.213611.116

Shen, W., J. Zhang, B. Geng, M. Qiu, M. Hu *et al.*, 2019 Establishment and application of a CRISPR-Cas12a assisted genome-editing system in *Zymomonas mobilis*. Microb. Cell Fact. 18: 162. https://doi.org/10.1186/s12934-019-1219-5

Silver, J. M., and N. R. Eaton, 1969 Functional blocks of the ad1 and ad2 mutants of *Saccharomyces cerevisiae*. Biochem. Biophys. Rep. 34: 301–305. https://doi.org/10.1016/0006-291X(69)90831-6

Stovicek, V., I. Borodina, and J. Forster, 2015 CRISPR-Cas system enables fast and simple genome editing of industrial *Saccaromyces cerevisiae* strains. Metab. Eng. Commun. 2: 13–22. https://doi.org/10.1016/j.meteno.2015.03.001

Van Rossum, G., and F. L. Drake, 2009 Python 3 References Manual. Scotts Valley CA: CreateSpace. https://www.python.org.

Wang, Y., Z. T. Zhang, S. O. Seo, P. Lynn, T. Lu *et al.*, 2016 Bacterial Genome Editing with CRISPR-Cas9: Deletion, Integration, Single Nucleotide Modification, and Desirable "Clean" Mutant Selection in *Clostridium beijerinckii* as an Example. ACS Synth. Biol. 5: 721–732. https://doi.org/10.1021/acssynbio.6b00060

Communicating editor: J. Berman