Bioinformatics, 34(1), 2018, 16–23
doi: 10.1093/bioinformatics/btx564
Advance Access Publication Date: 8 September 2017
Original Paper

OXFORD

Genome analysis

Enhanced guide-RNA design and targeting analysis for precise CRISPR genome editing of single and consortia of industrially relevant and non-model organisms

Brian J. Mendoza[1,2] and Cong T. Trinh[1,2,*]

[1]Department of Chemical and Biomolecular Engineering, University of Tennessee, Knoxville, TN 37996, USA and [2]Oak Ridge National Laboratory, Bioenergy Science Center (BESC), Oak Ridge, TN 37830, USA

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on March 20, 2017; revised on September 2, 2017; editorial decision on September 4, 2017; accepted on September 6, 2017

Abstract

Motivation: Genetic diversity of non-model organisms offers a repertoire of unique phenotypic features for exploration and cultivation for synthetic biology and metabolic engineering applications. To realize this enormous potential, it is critical to have an efficient genome editing tool for rapid strain engineering of these organisms to perform novel programmed functions.

Results: To accommodate the use of CRISPR/Cas systems for genome editing across organisms, we have developed a novel method, named CRISPR Associated Software for Pathway Engineering and Research (CASPER), for identifying on- and off-targets with enhanced predictability coupled with an analysis of non-unique (repeated) targets to assist in editing any organism with various endonucleases. Utilizing CASPER, we demonstrated a modest 2.4% and significant 30.2% improvement (*F*-test, $P < 0.05$) over the conventional methods for predicting on- and off-target activities, respectively. Further we used CASPER to develop novel applications in genome editing: multitargeting analysis (i.e. simultaneous multiple-site modification on a target genome with a sole guide-RNA requirement) and multispecies population analysis (i.e. guide-RNA design for genome editing across a consortium of organisms). Our analysis on a selection of industrially relevant organisms revealed a number of non-unique target sites associated with genes and transposable elements that can be used as potential sites for multitargeting. The analysis also identified shared and unshared targets that enable genome editing of single or multiple genomes in a consortium of interest. We envision CASPER as a useful platform to enhance the precise CRISPR genome editing for metabolic engineering and synthetic biology applications.

Availability and implementation: https://github.com/TrinhLab/CASPER.

Contact: ctrinh@utk.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Transforming biology into engineering practice has shaped the frontiers of synthetic biology and metabolic engineering (Connelly *et al.*, 2015; Trinh and Mendoza, 2016). This helps drive industrialization of biology with broad applications related to food, health, energy, and the environment from production of drugs to fight diseases to synthesis of renewable fuels to replace fossil fuels (Nielsen and Keasling, 2016). Genetic diversity of non-model organisms offers a repertoire of unique phenotypic features for exploration and cultivation for these applications (Jullesson *et al.*, 2015; Lee *et al.*, 2012).

To realize this enormous potential, it is critical to have an efficient genome editing tool for rapid strain engineering of these organisms to perform novel programmed functions. In the last several years, the CRISPR technology has emerged as a powerful genome editing tool for metabolic engineering and synthetic biology applications (Cong *et al.*, 2013; Farasat and Salis, 2016; Garst *et al.*, 2017; Gomaa *et al.*, 2014; Hsu *et al.*, 2014; Jakočiūnas *et al.*, 2015; Li *et al.*, 2013; Platt *et al.*, 2014; Qi *et al.*, 2012, 2013; Sander and Joung, 2014; Zeitoun *et al.*, 2015).

Using a CRISPR/Cas system for genome editing requires it to be both highly active and accurate. CRISPR tools that are used to design guide-RNAs (gRNAs) to form ribonucleoprotein (RNP) complexes for effective genome targeting focus on two types of evaluation. The first type is the prediction of 'on-target' activity of a selected gRNA on a target DNA (or RNA) sequence on the genome. This on-target activity reflects the ability of such a gRNA to successfully find and bind the Cas endonuclease to the target sequence. The second type is the prediction of 'off-target' activity of a selected gRNA, which pertains to the propensity of the RNP complex to interact with sequences on the genome similar to the target sequence.

Choosing gRNA sequences with high activity is paramount to all genome editing endeavors for metabolic engineering and synthetic biology applications (Feng *et al.*, 2015; Ryan *et al.*, 2014; Stovicek *et al.*, 2015). To predict which sequences lend themselves to efficient targeting, various algorithms have been developed based on high-throughput experimentation and validation (Doench *et al.*, 2014, 2016; Moreno-Mateos *et al.*, 2015). By identifying prominent nucleotide features in highly active gRNAs and their target sequences, scoring tables were generated and a gRNA sequence may be cross-referenced to predict its on-target activity. Past experimental evidence has revealed that the CRISPR RNP complex can reach cleavage efficiencies of at least 80% for single site modification (Jakočiūnas *et al.*, 2015). However, this efficiency might become problematic when applying a CRISPR/Cas system to modify multiple sites across the genome simultaneously using multiple gRNAs, known as multiplexing, for rapid strain engineering. For instance, the probability of achieving activity at 5 loci simultaneously drops to 32% (0.8^5). Therefore, developing new algorithms capable of accurately designing gRNA sequences with high activity is important for multiplex CRISPR implementation.

In addition to the on-target analysis, it is very critical to make sure that a gRNA design has limited activity at non-targeted sites. The foundation for such off-target analysis for the spCas9 (Cas9 derived from *Streptococcus pyogenes*) system in eukaryotes was established by Hsu *et al.* (2013) and later by Lin *et al.* (2014). Studies building on this foundation generally take two directions: one being the engineering or discovery of alternative Cas enzymes that exhibit lower off-target activity (Kleinstiver *et al.*, 2015), and the other being algorithms that enhance prediction capabilities (Lin *et al.*, 2014). Recently developed software platforms have implemented various algorithms to determine off-target sites for a gRNA design, based on either biological principles or training a model to fit experimental data (Doench *et al.*, 2016; Hendel *et al.*, 2015; Kleinstiver *et al.*, 2016; Labun *et al.*, 2016). A collection of these programs is detailed at https://omictools.com/crispr-cas9-category. Unfortunately, some potential off-target sequences are unable to be reconciled by these algorithms that do not account for newly discovered interactions between the target DNA and RNP. For instance, a study by Malina *et al.* (2015) shows the presence of PAMs (Protospacer Adjacent Motifs) across the sequence to be inhibitory to Cas9 cleavage. This underscores the need for additional experimental investigations into off-target effects and a robust algorithm to capture these effects for designing gRNAs, especially when applying CRISPR/Cas systems to non-model organisms.

Furthermore, the rapid discovery of novel, diverse Cas enzymes beyond spCas9 requires a flexible and robust algorithm that can accurately predict on- and off-target activities utilizing new CRISPR/Cas systems for genome editing. For instance, the discovery and use of another type of CRISPR class II endonuclease, *Acidaminococcus* sp. Cpf1 (asCpf1), has shown complementary activity to the Cas9 family with a T-rich PAM and 'sticky-end' generation (Zetsche *et al.*, 2015). Based on the crystal structure of the asCpf1–RNA–DNA heterotriplex, it is possible to infer important nucleotides within the CRISPR-RNA (crRNA) for catalytic efficiency, and thus develop accurate targeting algorithms for genome editing (Kleinstiver *et al.*, 2016; Yamano *et al.*, 2016).

While previous research has focused on identifying highly active and unique target sites using on- and off-target algorithms to assist in gRNA design, targeting non-unique (repeated) sites with CRISPR tools may lead to some interesting future directions for experimentation (Prykhozhij *et al.*, 2015). This strategy, called multitargeting, has powerful metabolic engineering and synthetic biology applications, but has not yet been fully explored. For instance, repeated sequences may serve as sites where a single gRNA could induce knockouts across multiple orthologs of a gene, reducing the amount of heterologous machinery required to achieve the same effect. Furthermore, engineering consortia of organisms has been a burgeoning field in the past decade (Andrianantoandro *et al.*, 2006). CRISPR tools can potentially investigate and modify these consortia, which may not be achievable by conventional genetic manipulations. Software with the ability to implement these applications can pave the way for novel metabolic engineering and synthetic biology applications using CRISPR/Cas systems.

In this study, we developed a novel method, named CRISPR Associated Software for Pathway Engineering and Research (CASPER) that implemented flexible algorithms to guide precise genome editing. Combining both experimental data and newly discovered biological principles, CASPER formulated an improved scoring method for enhanced prediction of on- and off-target activities. This prediction presents 'relative' activities of a gRNA design that depend solely on target sequences, and hence are independent of experimental conditions (e.g. growth rate, Cas9 concentrations, and DNA supercoiling). These 'relative' activities are different from 'absolute' activities of a gRNA design that are condition-specific and must be determined experimentally (Farasat and Salis, 2016; Hsu *et al.*, 2013). Further, CASPER expanded novel applications of CRISPR/Cas systems in genome editing including multitargeting analysis (i.e. simultaneous multiple-site modification on a target genome with single gRNA requirement) and multipopulation analysis (i.e. gRNA design for genome editing of a consortium of organisms).

2 Materials and methods

2.1 On-target activity formulation for CRISPR gRNA design

On-target activity is defined as the binding affinity and subsequent nuclease activity of a particular gRNA and endonuclease complex to its matching DNA (or RNA) target sequence. Many factors contributing to absolute on-target activities have been identified such as the state of DNA supercoils, genome size, Cas endonuclease concentration, and growth rate (Farasat and Salis, 2016). However, these factors have the same effect on every target within a genome and

therefore have no impact on the difference in relative activity between targets found on the same genome. Focusing on relative on-target activities enables us to formulate CASPER to analyze a large set of non-model organisms and alternative Cas endonucleases.

To predict the on-target activity of the complex, our developed CASPER method calculates the score $S_{C,\ P}$ (Equation 1) that is defined as follows:

$$S_{C,P} = \frac{S_C}{P} \tag{1}$$

where S_C is the CRISPRscan (Moreno-Mateos *et al.*, 2015) score and P is the penalty score. In Equation (1), $S_{C,P}$ has a value between 0 and 100 that is assigned to a gRNA seed sequence, with higher values indicating higher predicted activity. The CRISPRscan score S_C (Equation 2) is defined as follows:

$$S_{C,P} = \sum_{i=l-6}^{l+6} S_{C,i} \tag{2}$$

where l is the length of a seed sequence and PAM (see Supplementary Fig. S1 for indexing illustration), and $S_{C,\ i}$, is the score associated with the mono and dinucleotide features at position i (N_i or NN_i) of the seed sequence that has been experimentally determined to be relevant (Moreno-Mateos *et al.*, 2015). Each score, $S_{C,\ i}$, is the sum of the values of the features appearing at position i, and the sum of all positions' scores (S_C) is normalized with respect to the highest and lowest possible scores to obtain a normalized value from 0 to 100. The values of features at any given position can be seen in the CRISPRscan scoring table (Supplementary File S1).

The penalty score, P (Equation 3), is obtained by the combination of the PAM density score, s_{ij}, (Supplementary Table S1) and the score S_G (Equation 4):

$$P = \begin{cases} s_{ij}S_G & s_{ij} > 1 \\ 1 - S_G/5 & s_{ij} = 1 \end{cases} \tag{3}$$

The score s_{ij} in Equation (3) is determined by the number of PAMs present in a given seed sequence, where i is the number of PAMs in the forward sequence and j is the number of PAMs in the reverse complement of the sequence. The score s_{ij} can be looked up in the i^{th} column of the j^{th} row in Supplementary Table S1, which is derived from experimental data (Malina *et al.*, 2015). Due to the length of the spCas9 PAM (3) and the length of a seed sequence (20), $i + j < 8$. In Equation (3), the score S_G, used to reinforce the importance of guanines and adenines to the stability/instability of the gRNA respectively, is formulated to account for the nucleotide composition of the seed sequence irrespective of position.

$$S_G = \sum_{i=1}^{l-PAM} n_i \tag{4a}$$

$$n_i = \begin{cases} 1 & if\ N_i = G \\ 0.5 & if\ N_i = C \\ -0.1 & if\ N_i = A \end{cases} \tag{4b}$$

The values n_i were derived to maximize correlation between scores and on-target experimental data (Moreno-Mateos *et al.*, 2015) further described in Section 3. A visual summary of the process of obtaining an on-target score for a sequence is shown in Supplementary Figure S1.

2.2 Off-target formulation for CRISPR gRNA design

In contrast to the on-target activity assessment, the off-target activity is defined as the probability of a given gRNA sequence to interact with a non-matching sequence on the genome. Our developed CASPER method calculates the off-target score $S_{H,\ T,\ S}$ as follows:

$$S_{H,T,S} = \frac{\left(\sqrt{S_H} + S_T\right) R^2 S_S^6}{4} \tag{5}$$

where the off-target score $S_{H,T,S}$ lies between 0 and 1, with a higher value indicating higher probability of off-target activity, and is determined by combining four subscores S_H, S_T, S_S, and R discussed below. The appropriate arrangement of these scores in Equation (5) was determined by employing a genetic algorithm using the Pearson's coefficient between the output scores and experimental cleavage efficiency (Hsu *et al.*, 2013) as the fitness function. A more detailed description of the algorithm is presented in Section 3.

The subscore S_H accounts for the types of nucleotide mismatches and their location on the seed sequence, and is defined as follows:

$$S_H = \prod_{i=1}^{20} M_{ij} \tag{6}$$

where M_{ij} is the element of the Hsu matrix derived from experimental data gathered in Hsu *et al.* (2013) (Supplementary Table S2), i represents the index of the mismatch, and j corresponds to the identity of the mismatch (e.g. C with A).

The subscore S_T is derived from the inverse relationship of the proximity of the mismatch to the PAM:

$$S_T = \frac{3.5477 - \sum_{i=1}^{20} \frac{1}{i}}{3.5477} \tag{7}$$

where S_T is valued from 0 to 1 with a higher score indicating a higher probability of off-target activity, and i is the index of the mismatch. This score was motivated by previous studies showing the farther a mismatch is from the PAM site, the less likely it is to interfere with activity (Anderson *et al.*, 2015). As a result, sequences with PAM distal mismatches are more likely to be sites of off-target activity. Summing the values for a mismatch at every location across the seed sequence gives a value of 3.5477, thus, this number is used to normalize S_T as formulated in Equation (7).

The subscore S_S also captures an inverse relationship of the proximity of the mismatch to the PAM but is formulated using a stepped scale:

$$S_S = 1 - \sum_{i=1}^{20} a_i \tag{8a}$$

$$a_i = \begin{cases} 0.1 & i \le 6 \\ 0.05 & 6 < i \le 12 \\ 0.0125 & i > 12 \end{cases} \tag{8b}$$

where i is the index of the mismatch, and a_i is defined by a step function. The step sizes were derived to agree with the previous experimental report where mismatches in regions closer to the PAM are more detrimental to activity (Hsu *et al.*, 2013).

Using the knowledge that some gRNAs are more stable/active than others, the on-target activity scores of the target sequence and the off-target sequence are assembled into a ratio, R:

$$R = \frac{S_{C,P}^{off}}{S_{C,P}^{on}} \quad (9)$$

where $S_{C,P}{}^{\text{off}}$ and $S_{C,P}{}^{\text{on}}$ in Equation (9) are the on-target scores for the target DNA sequences appearing at the undesirable and desirable sites of the genome, respectively. A ratio of $R < 1$ signifies that the sequence of interest (SOI) is more active compared to the potential off-target site, meaning less likelihood of the off-target site being hit compared to the SOI. A ratio $R > 1$ signifies the reverse, i.e. a highly active off-target site that has an increased likelihood of being hit compared to the SOI. The visual representation of gathering all these scores and combining them for a given two sequence comparison can be seen in Supplementary Figure S1.

2.3 Multitargeting analysis

Multitargeting is the process of editing multiple sites simultaneously across the genome of an organism using a single gRNA. Instead of removing repeated gRNA target sequences, our developed CASPER method stores the data of repeated sequences for further analysis. To generate data of the non-unique seed sequences, CASPER references the set of sequences that appear more than once to a genome annotation file generated by either GenBank (http://www.ncbi.nlm.nih.gov/genbank/) or KEGG (genome.jp/kegg/). These sequences are then sorted based on the number of times they appear in an annotated region. Further analysis can be performed by investigating individual sequences in order to design gRNAs to target these repeated sequences.

2.4 Multipopulation analysis

To analyze multiple genomes simultaneously for use in the study and genetic modification of a consortium, CASPER collects the data by identifying targets across a collection of genomes and performs on- and off-target activity analysis against the entire metagenome. This enables precise gRNA design for accurate genome editing within a consortium. To deal with polyploid organisms, the same algorithm can apply. Sequences shared across genomes can also be identified, thereby applying the concept of multitargeting to populations.

3 Results and discussion

3.1 Development of CASPER algorithm for enhanced on-target activity prediction

In order to analyze sequences to be targeted by a gRNA/Cas complex, we developed an algorithm for CASPER to predict the gRNA design for the on-target activity (Equation 1). It was formulated to incorporate three guiding principles: (i) the CRISPRscan features experimentally identified to be present in highly active guide sequences (Moreno-Mateos *et al.*, 2015), (ii) the density of the PAM in question across the guide sequence (Malina *et al.*, 2015), and (iii) the propensity for guanines (and to a lesser extent cytosines) over adenines in the gRNA sequence which has been shown to be a factor in gRNA stability (Doench *et al.*, 2014; Wang *et al.*, 2014). To formulate a combination of these three principles, we combined the normalized CRISPRscan score, S_C (Equation 2) and divided it by a penalty score, P (Equation 3). The structure of P is determined by whether or not the number of PAMs in the sequence is inhibitory ($s_{ij} > 1$). If so, the value is multiplied by the score S_G; otherwise, it is disregarded and P purely accounts for the nucleotide content of the sequence (the S_G score). While the values of s_{ij} are derived directly from experimental data (Malina *et al.*, 2015), the score S_G needs to be constructed *de novo*.

Since previous studies showed guanines were favorable and adenines unfavorable to gRNA stability (Doench *et al.*, 2014; Wang *et al.*, 2014), we decided to value guanine most favorably (positive value) and adenine least favorably (negative value). The values assigned to guanine, adenine, and cytosine (Equation 4b) were obtained by varying their values between 0 and 1 (with a step of 0.1) and then evaluating which combination of values resulted in the optimal Pearson coefficient between experimental data from Moreno-Mateos *et al.* (2015) and the score, $S_{C,P}$.

We applied CASPER's on-target activity algorithm to predict the on-target activities of gRNAs from the experimental study by Moreno-Mateos *et al.* (2015). This study investigated indel frequency from a pool of 1280 gRNAs in zebrafish single cell embryos at 9-h post-fertilization. The CASPER method provided a minor 2.4% improvement in the R^2 value as compared to the CRISPRscan model (Fig. 1A) by taking into account PAM density and position-independent nucleotide content. Such a minor variation underscores the robustness of the CRISPRscan model. This improvement is statistically insignificant from the CRISPRscan model as determined by an F-test between the two datasets ($\alpha = 0.05$). However, CASPER successfully identified three of the four largest CRISPRscan score outliers in an experimental dataset of 25 gRNAs (Fig. 1A).

CASPER's on-target activity algorithm can be refined with subsequent datasets or further studies on spCas9's underlying mechanism of action. Additional on-target datasets generated by future studies will allow CASPER to develop parameters to assess gRNA activity in novel CRISPR/Cas systems (i.e. diverse organisms and Cas enzymes). For instance, CASPER's on-target activity algorithm can be applied directly to asCpf1 as soon as experimental data of on-target activity becomes available. In the meantime, to evaluate asCpf1 and other non-canonical endonucleases, CASPER focuses on the principles governing PAM density and guanine/adenine content to predict on-target activity.

3.2 Development of CASPER algorithm for enhanced off-target activity prediction

The objective of off-target analysis is to determine the sequences most likely to exhibit activity despite not being an identical match to the designed gRNA. Many factors contribute to off-target activity, including

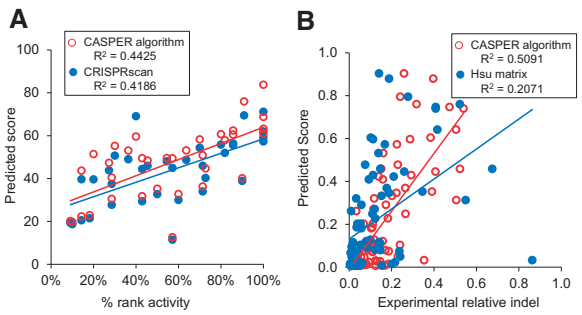

Fig. 1. (A) Comparison of CASPER on-target algorithm with CRISPRscan against experimental data obtained from Moreno-Mateos *et al.* (2015). The % rank activity (on-target) is obtained by ranking the activity of a sequence with respect to the rest of other sequences in the sample set. **(B)** Comparison of CASPER off-target algorithm with Hsu matrix method against experimental data obtained from Hsu *et al.* (2013) (off-target). The experimental relative indel (off-target) is determined by dividing the absolute indel % of the off-target site by the indel % of the targeted site. For example, a gRNA that has a 25.0% indel at the targeted site and 1.0% at the off-target site gives a 4% (or 0.04) relative indel

the number and location of mismatches, the number of times the PAM appears across the genome, and the concentration of the endonuclease. Only those variables relating to sequence identity are considered here, as other factors are considered to be either relatively constant (e.g. Cas9 concentration) or entirely unpredictable (e.g. state of DNA supercoils) when comparing one sequence relative to another. To perform such an analysis, an algorithm was developed for CASPER to compare all potential target sites in the genome to the SOI. For each comparison, mismatches between the two sequences are identified. If the total number of mismatches across the two sequences exceeds four, then the pairing is given a score of 0, signifying there should be no appreciable activity at the off-target sequence in question. The mismatches are then scored according to Equations (6), (7), and (8a). In addition, the on-target scores for the two sequences are obtained and the ratio (Equation 9) is also incorporated into a final score, $S_{H,T,S}$ (Equation 5). This scoring method improved prediction capabilities by 30.2% in the R^2 value (Fig. 1B) over the canonical study (Hsu *et al.*, 2013).

The enhanced correlation to experimental values could be attributed to the combination of incorporating experimental data (subscore S_H) (Hsu *et al.*, 2013) and guiding principles derived from the CRISPR RNP complex's mechanisms of action (subscores S_T, S_S, R). A genetic algorithm was employed to determine the optimum arrangement of the subscores (S_H, S_T, S_S, and R) to create a single final score, $S_{H,\ T,\ S}$, and to determine if any subscore was redundant. The result from the combinations of these scores were compared to experimental values given in Hsu *et al.* (2013) and the resulting correlation (R^2) was used as the fitness parameter for the genetic algorithm. The algorithm was given free range over modifying the coefficients and exponents for each of the subscores, as well as the general form of the equation, i.e. whether the subscores were added/subtracted or multiplied/divided to each other. The algorithm was initialized with 100 possibilities and the fittest 'parents' were chosen for crossover to create a new generation. Each generation spawned 100 children. By running the algorithm for 1000 generations, the final format present in Equation (5) was achieved. To confirm the algorithm was not over trained on the experimental dataset provided, a separate set of data (Hsu *et al.*, 2013) was used for determining the correlation of the algorithm's output with experimental off-target activity data (Fig. 1B).

Off-target guiding design principles of asCpf1 are similar to those of Cas9 in that the PAM proximal nucleotides are important for binding and enzyme activity, particularly the first eight (Zetsche *et al.*, 2015). Due to the lack of off-target experimental data at the time of this study, asCpf1 off-target identification was simplified to the use of Equations (7–9) as off-target data to generate a matrix for Equation (6). In addition, using the recently solved crystal structure of the asCpf1–RNA–DNA heterotriplex (Yamano *et al.*, 2016) enables us to exclude the 4 most PAM distal nucleotides of the 24-nucleotide gRNA sequence, as they are oriented away from pairing with the DNA when complexed with asCpf1. The parameters of Equation (8) were modified to give Equation (10):

$$S_S = 1 - \sum_{i=1}^{24} a_i \tag{10a}$$

$$a_i = \begin{cases} 0.1 & i \leq 8 \\ 0.0125 & 8 < i \leq 20 \\ 0 & i > 20 \end{cases} \tag{10b}$$

As in Equation (8b), the values for a_i in 10 b are calculated such that a completely mismatched sequence is given a score (S_S) of 0. This equation accounts for experimental observations that the most significant nucleotides in the seed sequence for asCpf1 are the eight nucleotides proximal to the PAM. In the absence of experimental data, to perform off-target activity analysis on non-canonical CRISPR endonucleases, such as asCpf1, CASPER subscores can be used independently to evaluate off-target activities.

The 30.2% increase in R^2 value was evaluated for statistical significance by an F-test using the scores from the Hsu matrix. It was confirmed that the CASPER algorithm represents a statistically significant improvement ($P < 0.05$).

The newly developed CASPER off-target algorithm can be quickly adapted to other systems seamlessly by training the algorithm against a set of experimental data and accounting for observed features associated with the biophysical properties of the CRISPR RNP complex. CASPER therefore provides a foundation for establishing *in silico* off-target analysis for a variety of organism and endonuclease combinations, especially when dealing with non-model organisms.

3.3 Development of multitargeting analysis

While sequences repeated in a genome have been traditionally discarded due to their inherent lack of specificity, we developed CASPER to exploit these sequences as potential gRNAs that are capable of targeting multiple sites across a genome with useful applications in synthetic biology and metabolic engineering.

To demonstrate CASPER for multitargeting analysis, we compiled data for 34 genomes of model and non-model organisms with three different endonucleases (spCas9, asCpf1, and spCas9-VRER) to gain perspective on the number of target sites that appear across each genome (Fig. 2). Genomes with a low GC content such as *Clostridium beijerinckii* (31.0%) and *C.saccharobutylicum* (29.4%) have a much greater number of target sites for asCpf1 (PAM: TTTN) than that of spCas9 (PAM: NGG). The spCas9-VRER variant (PAM: NGCG) in particular have only a couple thousand sequences appear for the mentioned organisms, a two order of magnitude difference compared to asCpf1. In addition, the greater PAM length of the VRER variant, which one would assume to appear less frequently than the canonical NGG, does indeed present less target sites across all the organisms investigated. These results show that some endonucleases are more useful than others depending on the genome and application. While this paper only presents a representative sample of non-model organisms, the algorithm is capable of analyzing any desired combination of organisms and endonucleases.

To understand where these repeated sequences appear, we mapped their locations on the annotated genomes. Figure 3 reveals that many of these sequences appeared in annotated regions, but a significant number also fell in unannotated ones, particularly sequences targeted by asCpf1 as shown by the significant number and height of black bars in Figure 3B and D. This discovery is valuable in that it helps reveal regions of the genome that may be related and subsequently probed through the precise CRISPR genome editing tool. Further analysis into non-unique targets in unannotated regions revealed a significant number of sequences in regions of completely unknown function. Sequences appearing in unannotated regions are particularly useful for two reasons. First, designing gRNAs to target sequences that appear in unannotated regions may reveal information about whether these regions are functional. Second, these sequences can be used for inserting multiple copies of genetic cargo with a reduced risk of unintentional disruption of cellular function. With multitargeting analysis, CASPER facilitates the

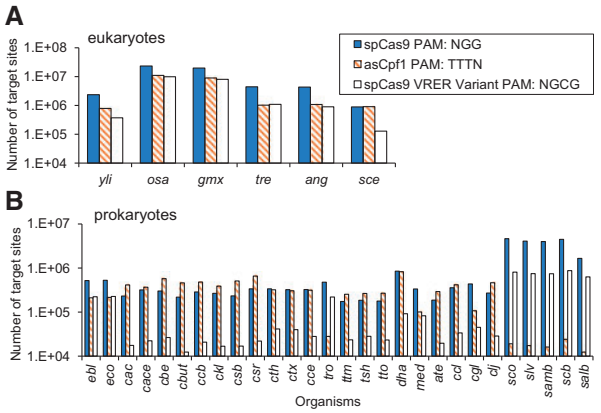

Fig. 2. Appearances of PAM motifs for three common Cas endonucleases (spCas9, asCpf1, and spCas9-VRER variant) across a select number of non-model organisms including **(A)** eukaryotes and **(B)** prokaryotes. Organism abbreviations: *yli: Yarrowia lipolytica, ang: Aspergillus niger, sce: Saccharomyces cerevisiae gmx: Glycine max (soybean) osa: Oryza sativa Japonica (rice) tre: Trichoderma reesei, cac: Clostridium acetobutylicum, cace: Clostridium aceticum, cbe: Clostridium beijerinckii, cbut: Clostridium butylicum, ccb: Clostridium cellulovorans, ckl: Clostridium kluyveri, csb: Clostridium saccharobutylicum, csr: Clostridium saccharoperbuytlacetonium, cth: Clostridium thermocellum (ATCC 27405), ctx: Clostridium thermocellum (DSM 1313), cce: Clostridium cellulolyticum, tro: Thermomicrobium roseum, ttm: Thermoanaerobacterium thermosaccharolyticum (DSM 571), tsh: Thermoanerobacterium saccharolyticum, tto: Thermoanaerobacterium thermosaccharolyticum (M0795), dha: Debaryomyces hanseii, med: Megasphaera elsdenii, ate: Caldicellulosiruptor bescii, ccl: Clostridium clariflavum, cgl: Cornyebacterium glutamicum, clj: Clostridium ljungdahlii, sco: Streptomyces coelicolor, slv: Streptomyces lividans, samb: Streptomyces ambofaciens, scb: Streptomyces scabiei, salb: Streptomyces albus*

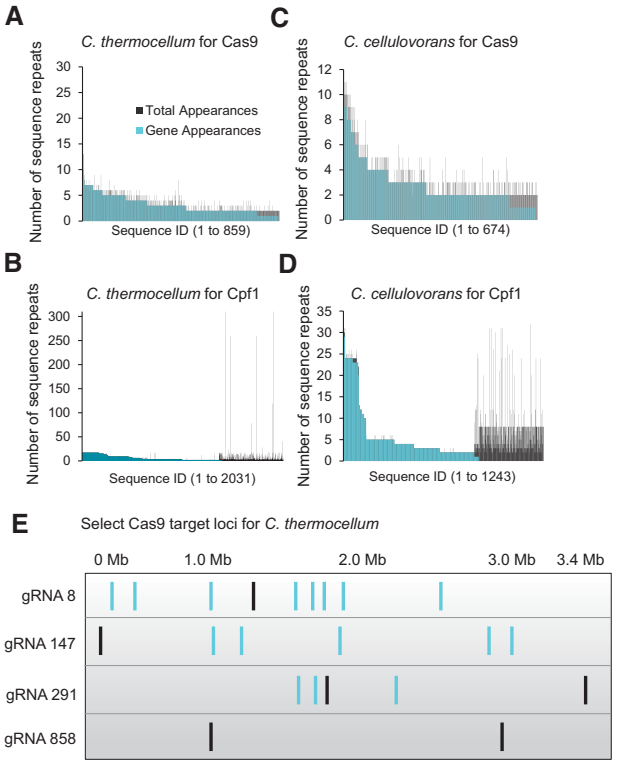

Fig. 3. Multitargeting analysis for each repeated sequence plotted by the number of times it appears within a gene (blue-green) and throughout the genome (black) for **(A)** *C.thermocellum* using spCas9, **(B)** *C.thermocellum* using asCpf1, **(C)** *C.cellulovorans* using spCas9, and **(D)** *C.cellulovorans* using asCpf1. Parentheses indicate the number of repeated sequences. **(E)** The distribution across the genome of *C.thermocellum* using a set of 4 representative non-unique sequences, including gRNA8: ACGTCAAACCTTTGGGCATT, gRNA147: GCAGAGCTGGATGAACATCT, gRNA291: TCCGTCGGTATCGGCTCCTC, and gRNA858: GACGGAGTCGGTTCATCAGA. Blue-green lines indicate appearance within a gene sequence while black lines represent appearances in un-annotated regions (Color version of this figure is available at *Bioinformatics* online.)

investigation and manipulation of unannotated regions with the CRISPR/Cas system for rapid genome editing.

To discover cellular processes that lend themselves to multitargeting, we looked further into the identity of the annotated regions in which these sequences were appearing. A cursory examination of the repeated sequences across the genomes revealed that transposable elements appear frequently across all organisms and can provide a powerful platform for integration of multiple copies of genes or entire operons (Fig. 4). Across the species investigated, between half and two-thirds of the non-unique sequences found on annotated regions were located on a transposon related feature. The promise of harnessing transposons for genome manipulation has been well documented by the utility of the Sleeping Beauty transposon system (e.g. SB10) (Geurts *et al.*, 2003; Ivics *et al.*, 1997). In addition to transposable elements, regions labeled as hypothetical proteins are also common sites to find repeated sequences (Fig. 4). Targets within hypothetical protein regions may be useful in a similar manner to those appearing in unannotated regions because the function of these cryptic regions can be systematically investigated.

Of particular interest are also the repeated sequences that appear on gene loci. Figure 4 details the common motifs targeted by non-unique sequences across the genomes of *C.thermocellum*, *C.aceticum*, and *C.kluyveri*. Interestingly, elements such as the cellulosome anchoring protein of *C.thermocellum*, a defining feature of the organism, is targeted 197 times by repeated sequences, opening the possibility of interrogating this structure with a select number of gRNAs. In *S.cerevisiae*, hexose transporters (416 times) and heat-shock proteins (111 times) are cellular processes that also may be targeted simultaneously by a single gRNA and are of particular interest for metabolic engineers looking to probe the sugar metabolism and environmental sensitivity and adaptation in this organism. In general, CASPER is capable of performing multitargeting analysis in any organism and with any endonuclease desired.

3.4 Development of multipopulation analysis

When editing a consortium of organisms with a CRISPR/Cas system, one must be mindful of the existence of similar or even shared sequences among genomes. Horizontal gene transfer of a plasmid containing the CRISPR/Cas system can exhibit activity in other species' genomes within a consortium. Additionally, direct transformation of the RNP into a consortium may result in unintended off-target activity in multiple species. It is therefore important to screen all genomes in a consortium for potential off-targets. We have developed the CASPER off-target algorithm to check for off-target sites across genomes in the consortium, thus minimizing unintended activity not just within the targeted organism but the consortium as a whole. In addition, CASPER can identify repeated sequences across the genomes that may lend themselves to multitargeting. This enables the identification of potential sites where multiple organisms in the consortium can be edited with the same gRNA.

To examine more closely the opportunities for multitargeting in a consortium, we ran CASPER against pairs of *Clostridial* species. Figure 5A and B details the number of sequences available for such targeting between the pairing of *C.thermocellum* and a selection of

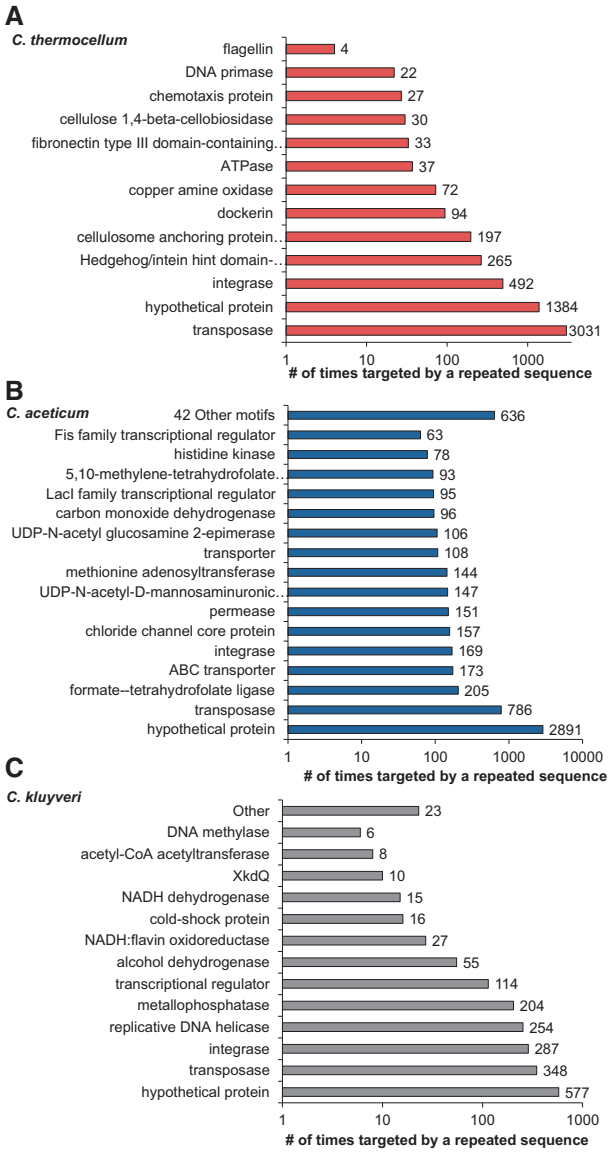

Fig. 4. Annotated cellular processes targeted by repeated sequences for **(A)** *C.thermocellum,* **(B)** *C.acetobutylicum,* and **(C)** *C.kluyveri*

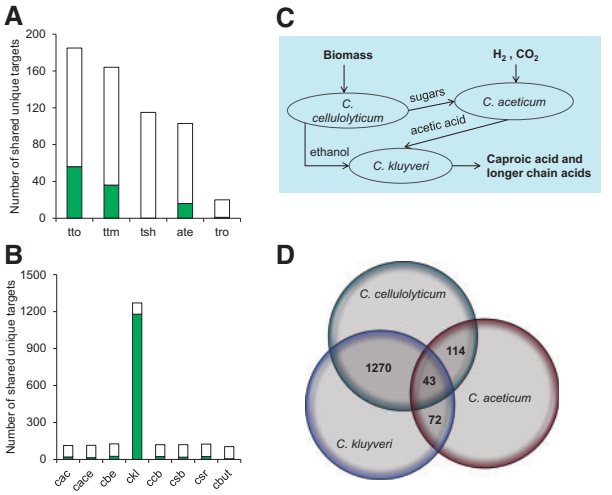

Fig. 5. Comparison of the number of shared targets between **(A)** *C.thermocellum* and other thermophiles and **(B)** *C.cellulolyticum* and other mesophiles. The colored portion of the bar represents the number of sequences that appear in both genomes but are unique within each genome. White portion of the bar represents the number of sequences that appear in both genomes and is repeated within one or both genomes being compared. **(C)** A hypothetical synthetic consortium of three different *Clostridial* species with synergistic metabolisms. The Venn diagram represents a three-way comparison of these organisms with the numbers of sequences shared among the genomes shown at the intersections (Color version of this figure is available at *Bioinformatics* online.)

thermophilic species, as well as the pairing of *C.cellulolyticum* with other mesophiles. Additionally, CASPER is capable of performing an analysis on multiple organisms to identify repeated sequences shared across them. For instance, we applied CASPER to identify repeated sites across *C.kluyveri, C.cellulolyticum*, and *C.aceticum* (Fig. 5C and D). This consortium was designed to utilize either biomass or a hydrogen/carbon dioxide feed for production of industrially relevant long-chain organic acids. This example demonstrates how CASPER can be used to identify off-targets and multitargeting analysis for genome editing of a consortium.

The number of sequences that may be used to target multiple organisms is quite small when compared to the size of the genomes. Thus, we postulate that the usefulness of multitargeting across organisms lies in niche insertions or deletions, such as those targeting a gene with high sequence similarity across the organisms in question. The opportunities for such manipulation will vary drastically depending on the identity of a consortium. Further, off-target analysis for a consortium will prove crucial in preventing unintended activity. Overall, CASPER's algorithms can facilitate the use of CRISPR tools to edit target genomes within a heterogeneous consortium.

The multipopulation analysis framework can also be utilized to analyze polyploid organisms. If individual allele sequences are known, they can be uploaded as separate 'genomes' and then compared via multitargeting on- and/or off-target analyses. This enables the selection of target sequences both unique and shared among alleles, facilitating partial or complete allelic editing with CRISPR/Cas.

4 Conclusion

The development of CASPER's flexible algorithms for analyzing on- and off-target activity in any organism with any Cas enzyme broadens the utility of CRISPR tools for genome editing of industrially relevant and non-model organisms. CASPER's multitargeting analysis facilitates simultaneous genetic manipulation of multiple loci using a single gRNA with novel potential applications including the investigation of large complex systems (e.g. the cellulosome of *C.thermocellum*) and unannotated genome regions. Further, CASPER's multipopulation analysis provides the ability to investigate microbial consortia and apply the CRISPR tools to perform genetic manipulations on single or multiple organisms within the consortia. We envision CASPER will assist the progress of CRISPR genome editing for metabolic engineering and synthetic biology applications.

Acknowledgements

The authors would like to thank Trinh lab's members for useful comments.

Funding

The research was financially supported in part by the lab start up fund and the Sustainability Energy and Education Research Center (SEERC) seed fund

at the University of Tennessee, Knoxville, the National Science Foundation grants (MCB #1553250, CBET #1511881, and CBET #1360867), a subcontract by the BioEnergy Science Center (BESC), a U.S. Department of Energy Bioenergy Research Center funded by the Office of Biological and Environmental Research in the DOE Office of Science (DE-AC05-000R22725), and the DARPA YFA award (D17AP00023). The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the funding agencies.

Conflict of Interest: none declared.

References

Anderson,E.M. *et al.* (2015) Systematic analysis of CRISPR–Cas9 mismatch tolerance reveals low levels of off-target activity. *J. Biotechnol.*, **211**, 56–65.

Andrianantoandro,E. *et al.* (2006) Synthetic biology: new engineering rules for an emerging discipline. *Mol. Syst. Biol.*, **2**, Article 0028.

Cong,L. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.

Connelly,T. *et al.* (2015) *Industrialization of Biology: A Roadmap to Accelerate the Advanced Manufacturing of Chemicals*.

Doench,J.G. *et al.* (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.*, **32**, 1–13.

Doench,J.G. *et al.* (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotech.*, **34**, 184–191.

Farasat, I. and Salis, H.M. (2016) A biophysical model of CRISPR/Cas9 activity for rational design of genome editing and gene regulation. *PLoS Comput. Biol.*, **12**, e1004724.

Feng,X. *et al.* (2015) Metabolic engineering of *Saccharomyces cerevisiae* to improve 1-hexadecanol production. *Metab. Eng.*, **27**, 10–19.

Garst,A.D. *et al.* (2017) Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. *Nat. Biotechol*, **35**, 48–55.

Geurts,A.M. *et al.* (2003) Gene transfer into genomes of human cells by the sleeping beauty transposon system. *Mol. Ther.*, **8**, 108–117.

Gomaa,A.A. *et al.* (2014) Programmable removal of bacterial strains by use of genome-targeting CRISPR-Cas systems. *MBio*, **5**, e00928–e00913.

Hendel,A. *et al.* (2015) Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. *Nat. Biotechnol.*, **33**, 985–989.

Hsu,P.D. *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.

Hsu,P.D. *et al.* (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, **157**, 1262–1278.

Ivics,Z. *et al.* (1997) Molecular reconstruction of sleeping beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell*, **91**, 501–510.

Jakočiūnas,T. *et al.* (2015) Multiplex metabolic pathway engineering using CRISPR/Cas9 in *Saccharomyces cerevisiae*. *Metab. Eng.*, **28**, 1–10.

Jullesson,D. *et al.* (2015) Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals. *Biotechnol. Adv.*, **33**, 1395–1402.

Kleinstiver,B.P. *et al.* (2015) Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*, **523**, 481–485.

Kleinstiver,B.P. *et al.* (2016) High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature*, **529**, 490–495.

Labun,K. *et al.* (2016) CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res.*, **44**, W272–W276.

Lee,J.W. *et al.* (2012) Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat. Chem. Biol.*, **8**, 536–546.

Li,J.-F. *et al.* (2013) Multiplex and homologous recombination-mediated genome editing in *Arabidopsis* and *Nicotiana benthamiana* using guide RNA and Cas9. *Nat. Biotechnol.*, **31**, 688–691.

Lin,Y. *et al.* (2014) CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.*, **42**, 7473–7485.

Malina,A. *et al.* (2015) PAM multiplicity marks genomic target sites as inhibitory to CRISPR-Cas9 editing. *Nat. Commun.*, **6**, 10124.

Moreno-Mateos,M.A. *et al.* (2015) CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods*, **12**, 982–988.

Nielsen, J. and Keasling, J.D. (2016) Engineering cellular metabolism. *Cell*, **164**, 1185–1197.

Platt,R.J. *et al.* (2014) CRISPR-Cas9 knockin mice for genome editing and cancer modeling. *Cell*, **159**, 440–455.

Prykhozhij,S.V. *et al.* (2015) CRISPR MultiTargeter: a web tool to find common and unique CRISPR single guide RNA targets in a set of similar sequences. *PLoS One*, **10**, e0119372.

Qi,L. *et al.* (2012) RNA processing enables predictable programming of gene expression. *Nat. Biotechnol.*, **30**, 1002–1006.

Qi,L.S. *et al.* (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, **152**, 1173–1183.

Ryan,O.W. *et al.* (2014) Selection of chromosomal DNA libraries using a multiplex CRISPR system. *Elife*, **3**, e03703.

Sander, J.D. and Joung, J.K. (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.*, **32**, 347–355.

Stovicek,V. *et al.* (2015) CRISPR–Cas system enables fast and simple genome editing of industrial *Saccharomyces cerevisiae* strains. *Metab. Eng. Commun.*, **2**, 13–22.

Trinh, C.T. and Mendoza, B. (2016) Modular cell design for rapid, efficient strain engineering toward industrialization of biology. *Curr. Opin. Chem. Eng.*, **14**, 18–25.

Wang,T. *et al.* (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, **343**, 80.

Yamano,T. *et al.* (2016) Crystal structure of Cpf1 in complex with guide RNA and target DNA. *Cell*, **165**, 949–962.

Zeitoun,R.I. *et al.* (2015) Multiplexed tracking of combinatorial genomic mutations in engineered cell populations. *Nat. Biotechnol.*, **33**, 631–637.

Zetsche,B. *et al.* (2015) Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell*, **163**, 759–771.