CASPER: An integrated software platform for rapid development of CRISPR tools

Brian Mendoza[1,§], Tanner Fry[2,§], David Dooley[1,§], Josh Herman[2], Cong T. Trinh[1*]

[1]Department of Chemical and Biomolecular Engineering, University of Tennessee, Knoxville, TN 37996, U.S.A.

[2]Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996, U.S.A.

[§] Equal contribution

*To whom correspondence should be addressed. Tel: +1 865-974-8121. Email: ctrinh@utk.edu

ABSTRACT

Both academic and enterprise software solutions exist for designing CRISPR targets. They offer advantages when designing gRNAs but often focus on a select number of model organisms. Those that offer a wide variety of organisms can be limited in support of alternative endonucleases and downstream analyses such as multitargeting and population analyses to interrogate a microbiome. To accommodate broad CRISPR utilization, we developed a flexible platform software CASPER (**C**RISPR **A**ssociated **S**oftware for **P**athway **En**gineering and **R**esearch) for gRNA generation and analysis in any organism and with any CRISPR-Cas system. CASPER combines traditional gRNA design tools with unique functions like multiple-Cas-type gRNA generation and evaluation of spacer redundancy in a single species or microbiome. The analyses have implications for strain-, species-, or genus-specific CRISPR diagnostic probe design and microbiome manipulation. The novel features of CASPER are packaged in a user-friendly interface to create a computational environment for researchers to streamline the utility of CRISPR-Cas systems.

BACKGROUND

The application of CRISPR-Cas systems has revolutionized biological research and industry alike.[1-6] The facile genome editing capabilities of CRISPR-Cas lie in its ability to specifically home in on a target sequence using a 20-40 base pair RNA sequence matched to a site on the genome/transcriptome. Most systems are further controlled with a protospacer adjacent motif (PAM) on either the 5' or 3' end of the sequence that interacts directly with the Cas protein complex and serves as a first checkpoint for substrate binding.[4, 7-9] The immense design space of targeting virtually any location on any genome with CRISPR-Cas machinery necessitates the development of rational sequence design tools to expedite experimental design.[10-12]

Central to the creation of gRNA sequences is the ability to score on- and off-target activities. Sequence similarities of gRNAs presumably determine whether an RNA-guided endonuclease can successfully identify a target site on a genome, and conversely, whether it will bind at off-target locations with high sequence similarity.[13] A variety of algorithms have worked off this premise;[14-16] however, further investigation into the fundamental biochemistry of spCas9's mechanism of action has illuminated other important features for analyzing on- and off-target binding and cleavage events.[8, 16, 17] For instance, PAM density and the formation of G-Q (guanine quadruplex) motifs have been shown to affect spCas9 activity.[18] Additionally, biomechanistic studies have provided insight into the kinetic and thermodynamic parameters associated with CRISPR-Cas enzymatic activity.[19-21] Finally, the increasing number of large assays to determine rules of activity are foundational to both identifying and confirming relevant features contributing to CRISPR-Cas activity.[15, 16, 22-28] Therefore, incorporating these features in the gRNA design algorithms helps select efficient guides for precise genome editing.

Currently, many such tools exist that employ either their own developed scoring algorithms or employ one or a series of previously developed algorithms to score identified guide sequences. While this scoring is a core element of any design tool, there exists a growing need for platforms with increased functionality as the applications of CRISPR-Cas diversify. The first wave of increased functionalities came with the advent of engineered Cas nucleases with unique PAMs and specificities, creating new applications such as CRISPRi/a (interference/activation) accompanied by experimentally validated rules for designing guides.[29-32] The rapid adoption of modified Cas proteins for base-editing and the associated rules for designing appropriate guide sequences further prompted the incorporation of new algorithms in existing software design platforms.[23, 33-35] Finally, use of CRISPR gene editing has rapidly expanded to interrogate microbiomes beyond single species.

This paper introduces CASPER (**C**RISPR **A**ssociated **S**oftware for **P**athway **E**ngineering and **R**esearch), an integrated software platform that contains features for designing CRISPR-Cas guides for diverse applications in non-model organisms and microbiomes. CASPER employs its own algorithms for on- and off-target analyses and has the unique capability of employing user-specified mismatch weighting parameters. In addition, CASPER provides end-users desired simplicity and flexibility in performing multitargeting analyses to exploit spacer redundancy in addition to conventional multiplexing analyses and designing gRNAs for editing either single or a consortium of organisms beyond select model species. Integration of the NCBI database in CASPER facilitates CRISPR-mediated manipulations for over 66,000 genomes with multiple Cas endonucleases and user-defined PAM specification options.

RESULTS

CASPER Infrastructure

To streamline analyses of CRISPR utility, we created a friendly graphical user interface (GUI) software, CASPER, to interrogate a single or consortia of species (Figure 1A). CASPER infrastructure is built using a modular framework. The current version is comprised of four modules, including New Genome, Target Finder, Multitargeting, and Population Analysis, which possess novel functionalities and can be expanded for future development (Figure 1B). Any analysis must start with New Genome to create a database of gRNA sequences stored in a *.cspr file for unique sequences and a *.db file for degenerate sequences (Figure 1B). In the Main Program, these files are used to perform the on-target and off-target analyses of gRNAs in the Target Finder module, identify the spacer redundancy in the Multitargeting module, and enable gRNA design to target specific species in a microbiome in the Population Analysis module (Figure 1B). With the inputs of organism(s), a type of endonuclease, and gene target(s) of interest, CASPER outputs the most likely active gRNA targets at the selected loci, with associated on- and off-target scores (Figure S5).

Organism and Endonuclease Support in the New Genome module

CASPER is designed to support genome editing across any organism through the New Genome module (Figure 2A). Using the NCBI tool within this module (Figure 2B), users may download and use any genomic sequence files (FASTA/FNA extension) and annotation files (GBFF extension) available in the GenBank and RefSeq collections of the vast NCBI database. Alternatively, users may upload a custom genome sequence themselves. Because of the genomic sequence flexibility, CASPER allows researchers to investigate the use of CRISPR/Cas systems to genetically modify communities of species and metagenomic assemblies.

In addition to multiple genome support, CASPER can generate guides for natural and engineered Cas9 orthologs[36] and other Cas endonucleases like AsCas12[11] in the built-in

endonuclease database within the New Genome module, and even novel, user-defined endonucleases using the "Define New Endonuclease" toolbox available in the Main Program. The latter functionality allows CASPER to be used for designing guide sequences that harness an organism's native CRISPR machinery. With the selection of annotated genome and endonuclease, running the New Genome module of CASPER will create databases of gRNA sequences for downstream analysis stored in the *.cspr and *.db files using the SeqFinder algorithm (Figure S1).

Guide RNA Selection and Library Generation in the Target Finder Module

The Target Finder module in the Main Program (Figure 3A) is designed to find gRNAs for a target gene(s) and assign them with relative on-target and off-target scores using our previously developed CASPERon and CASPERoff algorithms, respectively.[37] Within the Target Finder module, using either keyword or position-based searches, users can precisely find the desired region(s) of the genome they wish to target (Figure 3A). Once a gene or locus has been selected, users may continue with either bespoke or batch gRNA generation. This modular design enables users to perform multiplexing analyses to find multiple gRNAs targeting multiple loci. Seamless integration with annotation files makes CASPER's gene targeting fast and facile.

In addition to on- and off-targeting analyses, the View Targets GUI of CASPER provides several features to aid in custom gRNA selection, including gRNA visualization within the gene of interest, on-target and off-target scores, 5' motif filtering, and Co-Targeting analysis (Figure 3B). The Co-Targeting feature identifies sequences that have compatible PAM sequences for multiple endonucleases, thereby minimizing variables when performing experiments with multiple Cas systems to identify promising enzymes in the application of choice. When

working with Cas13 endonucleases or other Cas systems that have a permissive protospacer flanking sequence (PFS), users are encouraged to use CASPER with a more restrictive PAM requirement such that guides can be designed as inherently co-targeting. This not only saves memory by decreasing database file size but also allows near-identical guide sequences to be used with additional Cas systems. CASPER makes this process trivial with facile selection of multiple guides for export to a comma-separated (CSV) file.

Due to the usefulness and increasing practice of CRISPR screening techniques, as well as the hit-or-miss nature of gRNAs, it is extremely beneficial to facilitate the generation of large quantities of high-performing guides with minimal time expenditure. CASPER's Generate Library feature robustly addresses this issue by providing the end-user means to quickly create and export customizable libraries of gRNAs (Figure 3A, 3C). Users are given the option to select up to 10 gRNAs per locus, the region within the loci from which to pull them, minimum on-target and maximum off-target scores, and several other options. The library is then exported to a CSV file, making alteration or ordering of guides as effortless as possible. We envision this feature as a useful resource for the scientific community which will help bolster efficiency and productivity of various CRISPR assays.

Multitargeting and Population Analysis Modules

The manipulation of multiple genes and entire operons is the primary focus of many metabolic engineering and synthetic biology endeavours and CRISPR tools offer powerful resources to enable such activities.[1] As such, the CASPER user interface focuses on streamlining the design process for multi-gene disruption and/or insertion. While the ability to design gRNAs to target genes in model organisms is incredibly useful but hardly novel,[2]

CASPER distinguishes itself through two features that are particularly useful when dealing with non-model organisms: multitargeting and population analyses.

CASPER's Multitargeting module is a robust platform for targeting multiple sites across the genome with identical or similar gRNA sequences using the Multitargeting algorithm (Figure S2). While multiple gRNAs are used to target multiple loci in multiplexing, multitargeting utilizes one guide to target multiple loci. This analysis can be used to interrogate common motifs across a genome for manipulation or for integration of multiple copies of genes using a single gRNA template (Figure 4A). It can also identify conserved motifs in genome(s) representing potential sites for global cellular process manipulation (e.g., stress response, transposons).[37]

The rising interest in manipulation of microbiomes prompted the population analysis feature of CASPER (Figure 4B). In the Population Analysis module, users begin by selecting two or more organisms and clicking "Analyze Population." The Population Analysis algorithm (Figure S3) calculates conserved seeds at both pairwise and population levels. Shared motifs may be viewed in greater detail to elucidate their exact locations within each organism and which guides are best suited for targeting them. An overview of the targeting relationships of all species within the community is represented by a pairwise heatmap of shared genomic targets. By coupling population analysis with single gRNA analysis on the same *.cspr database files, users may find either degenerate gRNAs capable of editing an entire community at once or specific gRNAs that selectively edit a single species within a complex consortium. This approach has immediate application for CRISPR antimicrobial gRNA design,[38-40] wherein broad-spectrum gRNAs may be designed to target all strains of a pathogen of interest (POI), or strain-specific gRNAs could be found to eliminate a POI amidst

a highly similar commensal population. The gRNA design for populations opens yet another dimension in the scope of genome editing with CRISPR-Cas for a rapidly evolving field of study.

DISCUSSION

A variety of both academic and enterprise software solutions exist for designing CRISPR targets (Figure S4). They each offer advantages when designing gRNAs but very often focus on a select number of model organisms. Those that do offer a wide variety of organisms can be limited in their support of alternative endonucleases and further downstream analyses such as multitargeting and population analyses. In addition to being the only all-in-one GUI-based CRISPR design software for single and batch gRNA generation with on- and off-target scoring for any organism-endonuclease pair, CASPER possesses several novel features that distinguish it from other solutions, namely the ability to co-target genomic loci with compatible endonucleases and simultaneously target multiple sites in single or consortia of cell populations, a useful resource for metabolic engineering and synthetic biology applications.

The version of the software presented here is designed with inherent flexibility to accommodate both the input of users and expanding research into more refined studies of CRISPR-Cas activity across the organism/endonuclease design space. A defining feature of CASPER is its population analysis capability. We envision that as greater understanding of microbiome dynamics becomes available, this tool can evolve to serve the research community's needs, particularly in pairing appropriate design with experimental tools. In addition, a priority is put on maintaining compatibility with expanding databases, particularly for metagenomes, as this is critical for maintaining the platform as a cutting-edge tool.

In summary, CASPER exists at the intersection of powerful back-end algorithms and simple, intuitive graphical user interfaces, making it a potent CRISPR design tool for specialist and non-specialist alike. The addition of multitargeting and population analyses further differentiates CASPER from its contemporaries, opening the door to new and exciting analyses for multiple strains of a single viral or bacterial species, as well as bacterial communities and metagenomes. Immediately, these analyses have implications for strain-, species-, or even genus-specific CRISPR diagnostic probe design and microbiome manipulation. We expect CASPER to empower these and other applications of CRISPR biology, expanding their impact and utility.

MATERIALS AND METHODS

Algorithms. The current CASPER program contains 4 core algorithms: SeqFinder, CASPERon/off, Multitargeting, and Population Analysis. The SeqFinder algorithm used in the New Genome module processes a genomic sequence file to extract, sort, store, and retrieve gRNA sequences through indexing and string compression. The analysis of CASPER for a new genome runs at $O(N^2)$ speed, with N being the genome size. This algorithm has been optimized from our previous version[37] and adapted for the CASPER GUI to process large genomes' data and streamline users' input and output data. For instance, the employment of divide-and-conquer strategy in the SeqFinder algorithm to index large genomes resulted in a reduction of peak memory usage from 16+ GB down to 2 GB for human genomes, and a decrease in processing time from 360 to 30 minutes. The details of pseudocode for the SeqFinder algorithm can be found in Figure S1.[37]

The CASPERon and CASPERoff algorithms are core to the Target Finder module. These algorithms are derived from our previous versions[37] and adapted to fit the CASPER GUI to

10

streamline users' input and output data. Briefly, an evolutionary optimization is employed to enable multivariate regression on a large experimental dataset, taking into account i) the CRISPRscan features experimentally identified to be present in highly active guide sequences[41], ii) the density of the PAM in question across the guide sequence[18], 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 guide-RNA stability,[15, 42] and iv) nucleotide mismatches.[13, 43] A workflow of CASPER's target identification implementation is detailed in Figure S5.

Both the Multitargeting and Population Analysis algorithms use the output of the SeqFinder algorithm to process the unique and repeat sequences. These algorithms have been modified to fit the CASPER GUI, which manages the data inputs and outputs and downstream analyses. Details of the Multitargeting and Population Analysis algorithms can be found in Figures S2 and S3, respectively.

Computation and Programming languages. CASPER is designed as a standalone program capable of operating on Linux, Windows, and Mac OS X operating systems. CASPER provides an easy-to-use GUI for a suite of novel algorithms we previously developed for prediction of on- and off-target activities for gRNA design, as well as performing multitargeting and population analyses.[37] Algorithms for the creation of databases and identification of on- and off-targets have been implemented in C++. C++ was chosen due to the substantial memory management necessary for the creation of organism databases. The GUI is entirely implemented using the PyQt5 Python package, figures are generated with the popular Python graphing package, matplotlib, and front-end data handling/parsing is handled natively within Python.

CASPER Availability. CASPER can be downloaded as a standalone application via https://github.com/TrinhLab/CASPERapp. Documentation for installing CASPER and using the software is provided in Supplementary Material S2. In addition, video tutorials are available at the Trinh Lab YouTube channel (see Supplementary Material 2). The YouTube link is: https://www.youtube.com/channel/UC3qfnXxv4KqKlo1j_BSx-MQ.

DECLARATIONS

Conflicts of Interest

None declared.

Acknowledgements

We thank Trinh's lab members for proofreading, commenting, and testing CASPER.

Authors' Contributions

CT perceived and supervised the project. BM conceptually developed the back-end algorithms used in CASPER. TF did most of the coding (front-end and back-end) and assisted in the design of CASPER's GUIs. DD designed CASPER's overall workflow and GUI, managed the production of the software, and helped with coding. JH helped with coding the annotation file parsing/search and Generate Library features. BM, DD, and CT wrote the manuscript.

SUPPLEMENTARY DATA

Supplementary Material 1 includes Figures S1-S5

Figure S1: SeqFinder algorithm to generate the gRNA database

Figure S2: Multitargeting algorithm

Figure S3: Population Analysis algorithm

Figure S4: Comparison of available CRISPR design software

Figure S5: Workflow of CASPER

Supplementary Material 2. The CASPER manual

REFERENCES

1.   Stovicek V, Borodina I, Forster J. CRISPR–Cas system enables fast and simple genome editing of industrial Saccharomyces cerevisiae strains. *Metabolic Engineering Communications.* 2015;2:13-22. DOI: http://dx.doi.org/10.1016/j.meteno.2015.03.001

2.   Jakočiūnas T, Bonde I, Herrgård M et al. Multiplex metabolic pathway engineering using CRISPR/Cas9 in Saccharomyces cerevisiae. *Metabolic Engineering.* 2015:1-10. DOI: 10.1016/j.ymben.2015.01.008

3.   Jackson SA, McKenzie RE, Fagerlund RD et al. CRISPR-Cas: Adapting to change. *Science.* 2017;356:eaal5056. DOI: 10.1126/science.aal5056

4.   Chen JS, Doudna JA. The chemistry of Cas9 and its CRISPR colleagues. *Nature Reviews Chemistry.* 2017;1. DOI: 10.1038/s41570-017-0078

5.   Li X, Jousset A, de Boer W et al. Legacy of land use history determines reprogramming of plant physiology by soil microbiome. *The ISME Journal.* 2019;13:738-751. DOI: 10.1038/s41396-018-0300-0

6.   Pursey E, Sünderhauf D, Gaze WH et al. CRISPR-Cas antimicrobials: Challenges and future prospects. *PLoS Pathog.* 2018;14:e1006990. DOI: 10.1371/journal.ppat.1006990

7.   Jiang F, Doudna JA. CRISPR-Cas9 Structures and Mechanisms. *Annu Rev Biophys.* 2017;46:505-529. DOI: 10.1146/annurev-biophys-062215-010822

8.   Anders C, Niewoehner O, Duerst A et al. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature.* 2014;513:569. DOI: 10.1038/nature13579

9.   Makarova KS, Koonin EV. Annotation and Classification of CRISPR-Cas Systems In: *CRISPR: Methods and Protocols*. (Lundgren M, Charpentier E and Fineran PC; eds). Springer New York, New York, NY. 2015; pp. 47-75.

10.   Kleinstiver BP, Prew MS, Tsai SQ et al. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature.* 2015;523:481-485. DOI: 10.1038/nature14592

11.   Zetsche B, Gootenberg Jonathan S, Abudayyeh Omar O et al. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell.*163:759-771. DOI: 10.1016/j.cell.2015.09.038

12.   Kleinstiver BP, Prew MS, Tsai SQ et al. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature.* 2015;523:481-485. DOI: 10.1038/nature14592

13.   Hsu PD, Scott Da, Weinstein Ja et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature biotechnology.* 2013;31:827-832. DOI: 10.1038/nbt.2647

14.   Anderson EM, Haupt A, Schiel Ja et al. Systematic analysis of CRISPR–Cas9 mismatch tolerance reveals low levels of o1. Anderson, E. M. et al. Systematic analysis of CRISPR–Cas9 mismatch tole1. Anderson, E. M. et al. Systematic analysis of CRISPR–Cas9 mismatch tolerance reveals low levels of of. *Journal of Biotechnology.* 2015;211:56-65. DOI: 10.1016/j.jbiotec.2015.06.427

15.   Doench JG, Hartenian E, Graham DB et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol.* 2014;32:1-13. DOI: 10.1038/nbt.3026

16.   Doench JG, Fusi N, Sullender M et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotech.* 2016 (Research);34:184-191. DOI: 10.1038/nbt.3437

17.   Knight SC, Xie L, Deng W et al. Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science.* 2015;350:823-826. DOI: 10.1126/science.aac6572

18.   Malina A, Cameron CJF, Robert F et al. PAM multiplicity marks genomic target sites as inhibitory to CRISPR-Cas9 editing. *Nature Communications.* 2015;6:10124. DOI: 10.1038/ncomms10124

19.   Singh D, Sternberg SH, Fei J et al. Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. *Nat Commun.* 2016;7:12778. DOI: 10.1038/ncomms12778

20.   Klein M, Eslami-Mossallam B, Arroyo DG et al. Hybridization Kinetics Explains CRISPR-Cas Off-Targeting Rules. *Cell Rep.* 2018;22:1413-1423. DOI: 10.1016/j.celrep.2018.01.045

21.   Raper AT, Stephenson AA, Suo Z. Functional Insights Revealed by the Kinetic Mechanism of CRISPR/Cas9. *J Am Chem Soc.* 2018;140:2971-2984. DOI: 10.1021/jacs.7b13047

22.   Lin J, Wong KC. Off-target predictions in CRISPR-Cas9 gene editing using deep learning. *Bioinformatics.* 2018;34:i656-i663. DOI: 10.1093/bioinformatics/bty554

23.   Concordet J-P, Haeussler M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic acids research.* 2018;46:W242-W245. DOI: 10.1093/nar/gky354

24.   Moreno-Mateos MA, Vejnar CE, Beaudoin JD et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat Methods.* 2015;12:982-988. DOI: 10.1038/nmeth.3543

25.   Kleinstiver BP, Tsai SQ, Prew MS et al. Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nat Biotechnol.* 2016;34:869-874. DOI: 10.1038/nbt.3620

26.   Kleinstiver BP, Pattanayak V, Prew MS et al. High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature.* 2016 (Article);529:490-495. DOI: 10.1038/nature16526

27.   Tycko J, Barrera LA, Huston NC et al. Pairwise library screen systematically interrogates Staphylococcus aureus Cas9 specificity in human cells. *Nat Commun.* 2018;9:2962. DOI: 10.1038/s41467-018-05391-2

28.   Hsu PD, Scott DA, Weinstein JA et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol.* 2013;31:827-832. DOI: 10.1038/nbt.2647

29.   Larson MH, Gilbert LA, Wang X et al. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat Protoc.* 2013;8:2180-2196. DOI: 10.1038/nprot.2013.132

30.   Zalatan JG, Lee ME, Almeida R et al. Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. *Cell.* 2015;160:339-350. DOI: 10.1016/j.cell.2014.11.052

31.   Peters JM, Colavin A, Shi H et al. A Comprehensive, CRISPR-based Functional Analysis of Essential Genes in Bacteria. *Cell.* 2016;165:1493-1506. DOI: https://doi.org/10.1016/j.cell.2016.05.003

32.   Qi LS, Larson MH, Gilbert LA et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell.* 2013;152:1173-1183.

33.   Komor AC, Kim YB, Packer MS et al. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature.* 2016;533:420-424. DOI: 10.1038/nature17946

34.   Eid A, Alshareef S, Mahfouz MM. CRISPR base editors: genome editing without double-stranded breaks. *Biochem J.* 2018;475:1955-1964. DOI: 10.1042/BCJ20170793

35.   Labun K, Montague TG, Krause M et al. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.* 2019;47:W171-W174. DOI: 10.1093/nar/gkz365

36.   Esvelt KM, Mali P, Braff JL et al. Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nature methods.* 2013;10:1116-1121. DOI: 10.1038/nmeth.2681

37.   Mendoza BJ, Trinh CT. Enhanced guide-RNA design and targeting analysis for precise CRISPR genome editing of single and consortia of industrially relevant and non-model organisms. *Bioinformatics.* 2018;34:16-23. DOI: 10.1093/bioinformatics/btx564

38.   Bikard D, Euler CW, Jiang W et al. Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. *Nat Biotechnol.* 2014;32:1146-1150. DOI: 10.1038/nbt.3043

39.   Citorik RJ, Mimee M, Lu TK. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. *Nature biotechnology.* 2014;32:1141-1145.

40.   Beisel CL, Gomaa AA, Barrangou R. A CRISPR design for next-generation antimicrobials. *Genome biology.* 2014;15:1-4.

41.   Moreno-Mateos MA, Vejnar CE, Beaudoin J-D et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat Meth.* 2015 (Article);12:982-988. DOI: 10.1038/nmeth.3543

42.   Wang T, Wei JJ, Sabatini DM et al. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science.* 2014 (10.1126/science.1246981);343:80.

43.   Lin Y, Cradick TJ, Brown MT et al. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Research.* 2014;42:7473-7485. DOI: 10.1093/nar/gku402

FIGURE LEGENDS

Figure 1: (A) Graphical user interface of CASPER. **(B)** CASPER infrastructure. CASPER is written in C++ in the backend computation to process genome sequences and perform on- and off-targeting analyses. For the Frontend computation and GUI, python is used. CASPER contains four modules, including New Genome, Target Finder, Multitargeting and Population Analysis.

Figure 2: New Genome module. **(A)** The New Genome GUI creates a database of gRNA sequences (unique and repeat gRNA databases) for a new genome with a user's defined endonuclease. **(B)** The NCBI Tool GUI allows a user to download a genome(s) from NCBI.

Figure 3: Target Finder module. **(A)** The Target Finder GUI enables a user to identify gRNA libraries to target a specific DNA sequence(s) or genes. Multiple loci can be selected for multiplexing analysis. **(B)** The View Targets GUI analyzes on- and off-target activities of gRNA libraries and export the results. Co-targeting feature can also be performed to identify a guide that can be compatible with different endonucleases. **(C)** The Generate Library GUI selects a library of gRNAs that target a DNA sequence or gene filtered by their locations, on- and off-target activities.

Figure 4: **(A)** The Multitargeting module identifies a library of single guides targeting multi loci. **(B)** The Population Analysis module identifies a library of guides that target one or multiple loci within one or more species within a microbiome.

1 Supplementary Material 1

2

3 CASPER: An integrated software platform for rapid development of

4 CRISPR tools

5

6 Brian Mendoza[1,§], Tanner Fry[2,§], David Dooley[1,§], Josh Herman[2], Cong T. Trinh[1*]

7

8 [1]Department of Chemical and Biomolecular Engineering, University of Tennessee, Knoxville,

9 TN 37996, U.S.A.

10 [2]Department of Electrical Engineering and Computer Science, University of Tennessee,

11 Knoxville, TN 37996, U.S.A.

12

13 [§]Equal contribution

14 *To whom correspondence should be addressed. Tel: +1 865-974-8121. Email:

15 ctrinh@utk.edu

Figure S1: SeqFinder algorithm to generate the gRNA database

Algorithm 1 SeqFinder algorithm to generate the gRNA database

---

Input:

    Endonuclease (*string*), PAM (*string*), Multithreading (*boolean*), Direction-
ality (*boolean*, T = 5' PAM, F = 3' PAM), Generate repeats file (*boolean*),
Length of gRNA 5' of seed (*int*), Seed length (*int*), Length of gRNA 3' of
seed (*int*), Organism abbreviation (*string*), Output directory path (*string*),
CASPERinfo file path (*string*), FNA/FASTA file path (*string*), Organism
name (*string*), Notes (*string*)

1: open FNA/FASTA file
2: **while** file open **do**
3:     **for** chromosome in file: **do**
4:         append chromosome sequence (*string*) to vector
5:     **end for**
6: **end while**
7: create vector of possible PAMs   ▷ Ex. for NGG, create (GGG, AGG, etc.)
8: **for** chromosome in vector **do**   ▷ find targets
9:     **for** PAM in PAM vector **do**
10:         find PAM in chromosome   ▷ forward and reverse strands
11:         append compressed seed sequences (*int*) and locations (*int*) to vectors
12:         score full length sequences and store in vector
13:     **end for**
14:     store total number of seeds found (*int*)
15: **end for**
16: sort compressed seed vector   ▷ repeats seeds are neighbors
17: **for** compressed seed in vector **do**   ▷ process targets
18:     **if not** identical neighbor **then**   ▷ this is a unique seed
19:         append index (*int*) to unique index vector
20:     **else**   ▷ this is a repeated seed
21:         append index to repeats index vector
22:     **end if**
23: **end for**
24: **for** chromosome in chromosome vector **do**   ▷ write out unique data
25:     **for** index in unique index vector **do**
26:         find location of seed using index
27:         **if** seed within chromosome range **then**
28:             retrieve unique information at index
29:         **end if**
30:     **end for**
31:     write all uniques for chromosome to CSPR file
32: **end for**
33: **for** index in repeats index vector **do**   ▷ write out repeats data
34:     use index to group related repeats
35:     retrieve sequence information for grouped repeats
36:     write grouped repeats to SQL database
37: **end for**

Output: : CSPR database file (uniques) and SQL database file (repeats)

---

19   **Figure S2:** Multitargeting algorithm

20

Algorithm 2 Multitargeting algorithm

Input: SQL database (DB) file from New Genome

  1: Retrieve seeds and counts from DB file
  2: Generate seeds vs. repeats graph
  3: Generate repeats vs. ID number graph
  4: Populate repeated seed table
  5: Generate chromosome visualization
  6:     Retrieve data for selected seed
  7:     Create a rectangular space (static size) for each chromosome
  8:     Get location of repeats from chromosome and scale location to static size of rectangle

Output: : Repeated seed table and figures

21

22

Figure S3: Population Analysis algorithm

Algorithm 3 Population Analysis algorithm

Input: SQL database (DB) files from New Genome

1: Create a temporary DB file consisting of inner joins between each selected organisms' seeds
2: Generate pairwise conserved seeds heatmap
3:     Query temporary DB file for pairwise numbers of conserved seeds
4:     Plot data
5: Generate conserved seeds table
6:     Query temporary DB file for seeds conserved between all selected organisms
7:     Fill table.   ▷ If none found, leave table empty

Output: : Conserved seed table and heatmap

Figure S4: Comparison of available CRISPR design software

		Software Feature													
		Targeting					Organisms and Endonuclease Support				Database Annotation and Searching			Others		
Competitor		sgRNA library design	Multi-targeting	Populations targeting	Target-site ranking	Non-canonical off-targets	Multiple Cas orthologs	Organism database	Genomes available	Genome size limit	Search by Gene ID	Database integration	Search by sequence	Online/Offline	Programming knowledge	Type
	Benchling	Yes	No	No	Yes	Yes	Yes	User defined	Unlimited	N/A	Yes	ENSEMBL	Yes	Online	None	Enterprise
	CRISPick	Yes	No	No	Yes	No	Yes	Preloaded	Limited	N/A	Yes	Unknown	Yes	Online	None	Academic
	CasOFFinder	No	No	No	No	Yes	Yes	N/A	Unlimited	N/A	No	No	Yes	Online/Offline	None	Academic
	CASPER	Yes	Yes	Yes	Yes	Yes	Yes	NCBI/User-defined	Unlimited	N/A	Yes	NCBI Entrez	No	Online/Offline	None	Academic
	CHOPCHOP	No	No	No	Yes	No	Yes	Preloaded	Limited	N/A	Yes	ENSEMBL	Yes	Online/Offline	None	Academic
	CRISPOR	Yes	No	No	Yes	No	Yes	User-defined	Unlimited	2000 bp	No	No	Yes	Both	None	Academic
	CRISPRSCAN	No	No	No	Yes	No	Yes	Preloaded	Limited	N/A	Yes	ENSEMBL	Yes	Online	None	Academic
	E-CRISP	No	Yes	No	Yes	No	Yes (3' PAMs only)	Preloaded	Limited	N/A	Yes	ENSEMBL	Yes	Online	None	Academic
	FlashFry	No	No	No	Yes	No	Yes	User-defined	Unlimited	N/A	No	No	Yes	Offline	Extensive	Academic
	Guides	Yes	No	No	Yes	No	No	Preloaded	Limited	N/A	Yes	Unknown	No	Online	None	Academic
	Horizon Discovery	No	No	No	No	No	No	Preloaded	Limited	N/A	Yes	Unknown	No	Online	None	Enterprise
	IDT	No	No	No	Yes	No	No	User defined	Unlimited	1000 bp	No	No	Yes	Online	None	Enterprise
	Protospacer WB	No	Yes	No	Yes	Yes	No	User defined	Unlimited	N/A	Yes	No	Yes	Offline	None	Academic
	Synthego	No	No	No	Yes	No	No	Unknown	123,769	N/A	Yes	ENSEMBL	No	Online	None	Enterprise

27

28

29

30

31 **Figure S5:** Workflow of CASPER
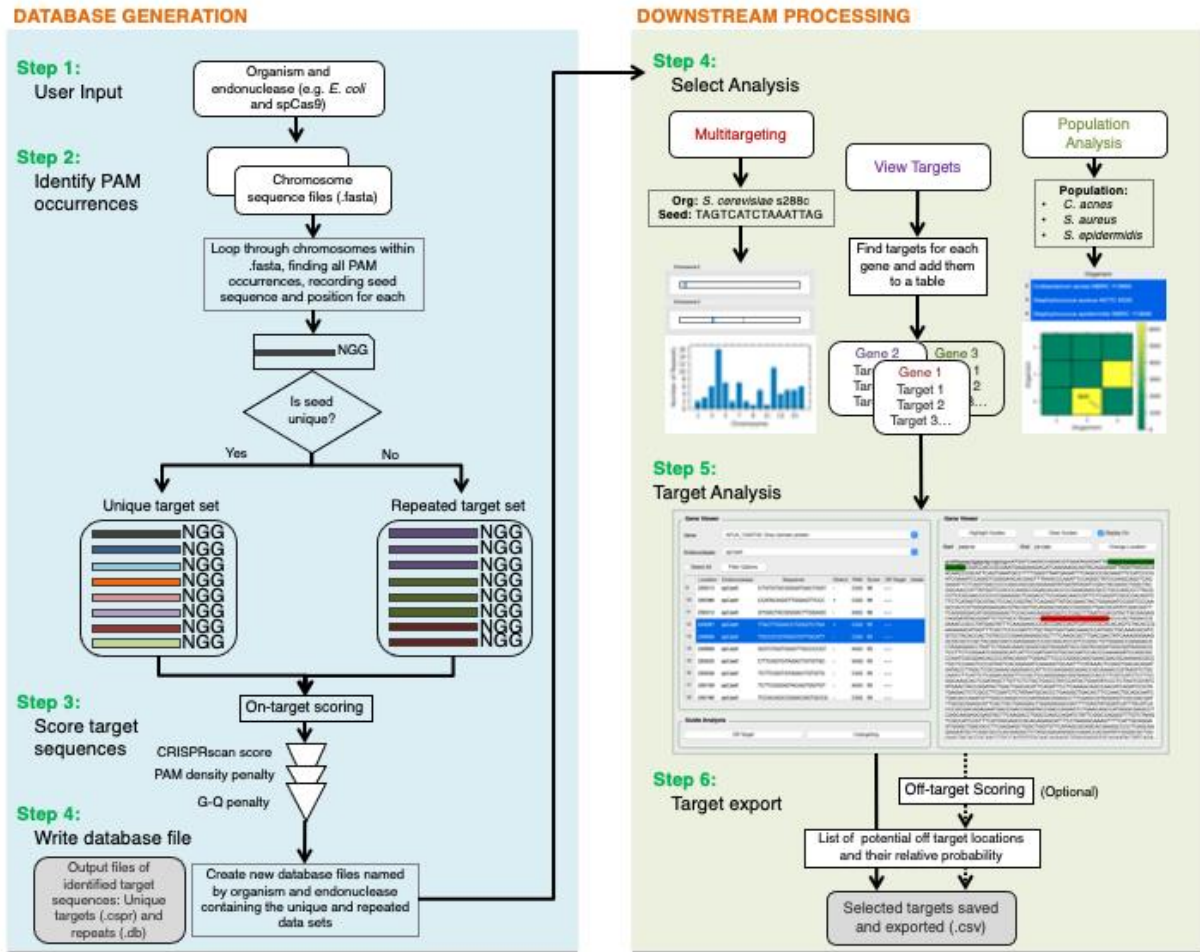
32

CASPER v2.0 Documentation

Trinh Lab
Department of Chemical and Biomolecular Engineering
University of Tennessee Knoxville
Last Updated: April 12, 2022

Contents

INTRODUCTION

Acknowledgements:

We thank the members of Trinh Lab for their suggestions, corrections, and testing of CASPER.

Getting Help:

Questions about CASPER may be directed to David Dooley at ddooley2@vols.utk.edu or posted on the Issues tab of the Github repository.

Background:

CRISPR-Cas systems have revolutionized biological research and industry alike. While these powerful systems have a broad range of applications, they are best known for their use as genome editors and nucleic acid detectors. Much of CRISPR-Cas' strength stems from its ability to specifically home in on a target sequence using a 20-40 base pair RNA sequence matched to a site on the genome/transcriptome. While there are certain restrictions on which sequences can be targeted, the design space for CRISPR-Cas guide RNAs (gRNAs) is magnificently immense. Therefore, development of rational sequence design tools has been necessary to expedite and improve experimental design.

Enter CASPER (**C**RISPR **A**ssociated **S**oftware for **P**athway **E**ngineering and **R**esearch), a tool which provides a flexible, robust platform for the analysis of non-model organisms and development of CRISPR-Cas tools capable of targeting a broad range of pathogens. CASPER is available for download at https://github.com/TrinhLab/CASPERapp.

CASPER Architecture:

CASPER is modular in nature and broken down into two distinct parts: back-end algorithms written in C++ and a graphical user interface (GUI) front-end written in Python (Figure 1).
The following algorithms are written in C++:

- **SeqFinder Algorithm**: Builds CSPR database, which includes on-target scoring. Source code available on Github.

- **Off-Target Algorithm**: Performs off-target scoring of gRNAs. Source code available on Github.
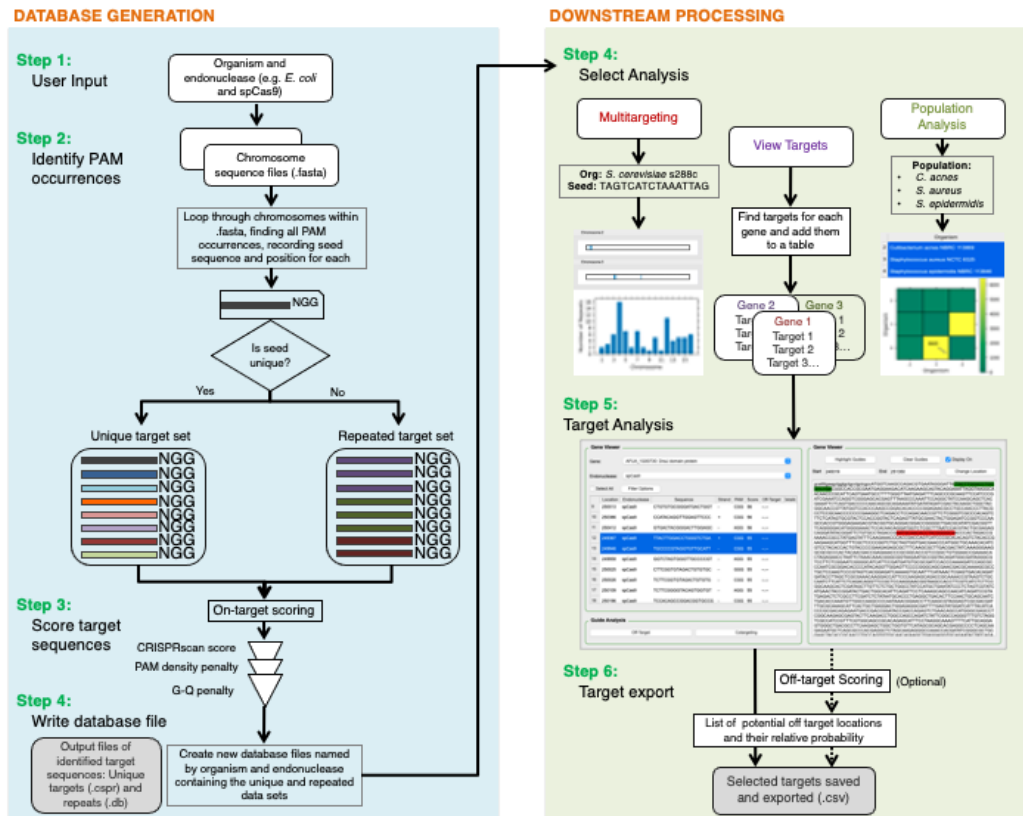
Figure 1: Overview of CASPER's architecture.

For more detailed information on how these algorithms were formulated, please refer to our previous publication here.

The GUI consists of four main modules, written in Python:

1. **New Genome**: Upload a genomic DNA file (FASTA/FNA) to generate a custom CSPR database for gRNA generation and analysis.

2. **Main Window**: Locate and evaluate gRNAs for a given gene(s) or location(s) of interest within a CSPR file (requires cognate annotation file).

3. **Multitargeting**: Analyze a CSPR file for repeated gRNA motifs within an organism.

4. **Population Analysis**: Analyze multiple CSPR files for conserved gRNA motifs across a population of organisms.

For detailed instructions on how to use CASPER, please refer to the **Navigating the Application** or **YouTube Tutorials** sections.

INSTALLATION

Installing from Github Releases:

To install CASPER from Github releases:

1. Navigate to the CASPER repository: https://github.com/TrinhLab/CASPERapp

2. Click on Releases on the right-hand side of the screen:
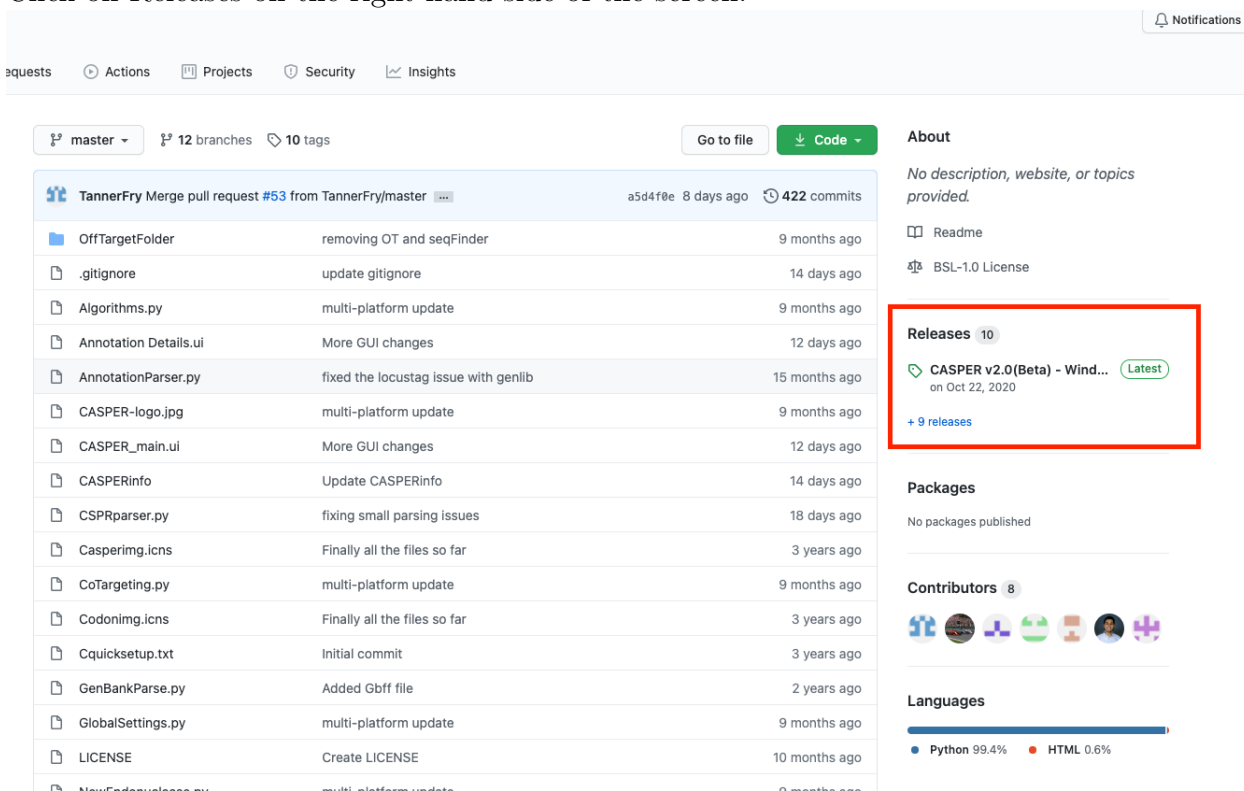
3. Scroll to the appropriate release for your operating system, click Assets, and download the installation package:

    - Windows
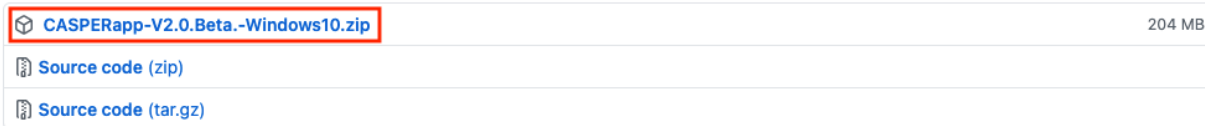
4. Follow the on-screen instructions for installing CASPER to your system. If on Mac or Linux, you must install the Source code file and follow the steps below, replacing Step 3 with the uncompressed Source code download.

Installing from Source Code:

To install CASPER from source code:

Note: this process is covered in detail in **Workflow 1**.

1. Ensure you have the following software packages installed and added to your path:

   - Python 3.8: Download here or install via Homebrew (*recommended*)
   - Git: Download here
   - Virtualenv: Download with pip:

     ```
     pip3 install virtualenv
     ```

   - For Windows Users, Visual Studio Build Tools: Download here

2. Open a terminal/command prompt window and `cd` to the directory you want to download the CASPER files into

3. Clone the repository by executing the command:

   ```
   git clone https://github.com/TrinhLab/CASPERapp.git
   ```

4. Create a virtualenv in the CASPERapp directory, specifying a 3.8 version of Python:

   ```
   virtualenv <name of environment> -p
   <path to python 3.8 executable>
   ```

5. Initialize the newly created virtualenv

   ```
   source <name of environment>/bin/activate
   ```

6. Upgrade pip and install dependencies from the same folder by executing the command:

   ```
   pip3 install --upgrade pip &&
   pip3 install -r requirements.txt
   ```

7. Run the program by executing the command:

   ```
   python3 main.py
   ```

   Note: before running CASPER, Mac users may have to give permission to Casper_Seq_Finder and OT executables. This can be done by following this tutorial.

8. Follow the instructions on screen to begin using CASPER. For more information on navigating CASPER, please refer to the **Navigating the Application** and **YouTube Tutorials** sections.

NAVIGATING THE APPLICATION

The CASPER GUI is composed of 4 primary modules, each of which carry their own features and functions:

- **New Genome**

- **Main Window**

- **Multitargeting**

- **Population Analysis**

This section will provide a basic understanding of what each of these 4 modules accomplishes, how to navigate them, and what inputs to provide them with. Detailed, step-by-step walkthroughs are deferred to the corresponding **YouTube Tutorials** at the end of this documentation.

The first window seen when starting CASPER is the launch window (Figure 2). The user starts by specifying a default directory for all CASPER-related files to be stored in. All files generated during CASPER analyses will be placed here, including CSPR database files (CSPR and DB files), genomic DNA files (FASTA/FNA), annotation files downloaded using the NCBI tool (GBFF), and exported gRNA files (CSV). If the user wishes to save files to a different location, they must either change the directory in the **Launch Window** or manually move the files outside of CASPER.

Next, the user is prompted to either proceed to the Main Window or New Genome. First-time users will have to generate a CSPR database in New Genome before continuing to Main Window. Once a CSPR file has been generated, the user may proceed directly to the Main Window when launching CASPER, if they so choose.
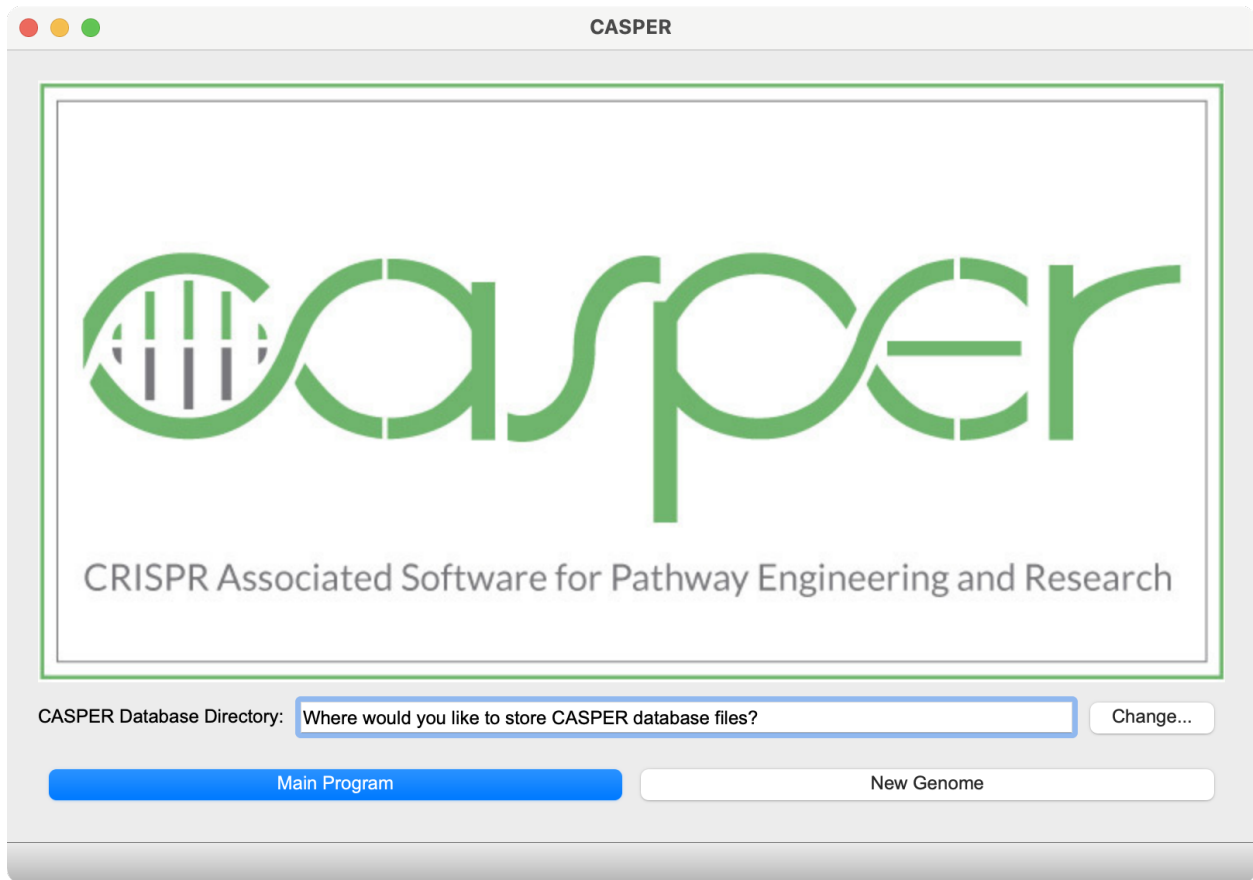
Figure 2: CASPER launch window.

<u>New Genome:</u>

New Genome is the module in which you may create new CSPR database files from an input FASTA/FNA file for the evaluation of unique and repeated gRNAs. It can be accessed via the **Launch Window** or the **Main Window** button: "Analyze New Genome". The window itself is essentially an entry form for the creation of a new CSPR file (Figure 3).

Figure 3: New Genome window.

The workflow is divided into 3 main steps:

1. Input Organism Info.

2. Select File and Add Job to Queue

3. Click "Run All".

The process of creating CSPR databases is covered in **Workflow 2** of the YouTube videos.

Note: for very large FASTA/FNA files (>1Gb), the runtime and memory requirements for CSPR database generation may become too great for the computer, depending on the robustness of hardware. In this case, we recommend breaking the file into multiple smaller files (~100Mb) of constituent parts (chromosomes/scaffolds/contigs), or only creating CSPR databases for the parts of the organism required for analysis.

Terminology:

- Organism Code: Manner in which CASPER will label its database files and references for the organism imported here. Ex: sau for *S. aureus* will yield files sau_<endonuclease>.cspr and sau_<endonuclease >_repeats.db

- Endonuclease: Cas enzyme to generate gRNAs for. Each enzyme has its own distinct targeting requirements (PAMs). Several have been preloaded into the program, but novel endonucleases may be added using the **New Endonuclease** feature.

- Seed Length: Length in nucleotides (nt) of the gRNA region in which binding is most crucial for Cas activity. CASPER's **SeqFinder Algorithm** uses this seed length and the endonuclease-specific PAM length to search for repeated gRNAs across the genome of interest.

- 5' Length: Length in nt of gRNA sequence upstream (5') of the seed sequence.

- 3' Length: Length in nt of gRNA sequence downstream (3') of the seed sequence.

- Multi-threading: Boolean (T/F) value for using multiple threads during the creation of CSPR files. It is highly recommended to leave multi-threading on, especially for larger genomes.

Main Window:

The Main Window (Figure 4) is opened by clicking "Main Window" in the **Launch Window**. It is the heart of CASPER, serving as the starting point for several different analyses and workflows, including:

- Monoplex/Multiplex Gene Targeting

- Multitargeting Analysis

- Population Analysis

- gRNA Library Generation
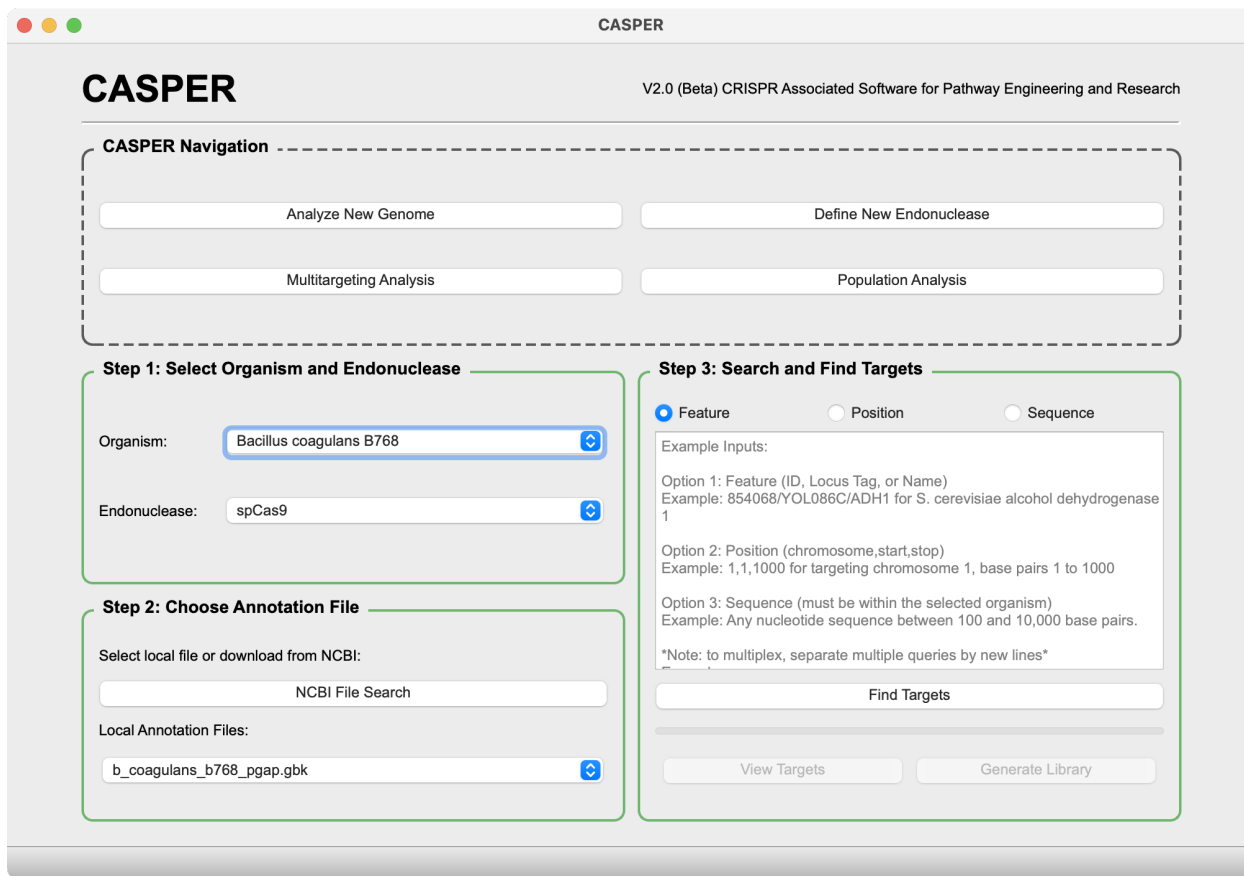
Figure 4: Main window.

The headlining feature of the CASPER software is the ability to do gene-based, location-baesd, or sequence-based gRNA generation using only an organism's genomic sequence file (FASTA/FNA) and annotation file (GBFF). It is important to note that the annotation file used for gene searches in this section must belong to *exactly* the same species and strain as the FNA/FASTA file used to create the CSPR file. Having an annotation file mismatched on even the strain level could eliminate all gene targeting functionality, since annotation files are arranged corresponding to the layout of their cognate genomic sequence files, which may be organized in a variety of ways (chromosomes/scaffolds/contigs).

The basic steps for generating gRNAs for a gene or locus of interest are:

1. Select Organism and Endonuclease
   - All previously generated CSPR files will automatically be populated from within the directory specified in the **Launch Window**.

2. Choose Annotation File
   - An annotation file may be downloaded manually using a web browser or by using the **NCBI Tool** provided in the Main Window/New Genome. Annotation files are automatically populated from within the directory specified in the **Launch Window**.

Note: the Feature and Sequence search options rely on GenBank-formatted annotation files, so if one is not available, users may only search using Position. To do this, simply select "None" from the list of Local Annotation Files and type the desired loci in the specified format (chrom,start,stop) with each new entry separated by a new line.

3. Search and View Targets
   - Searches can be in the form of either gene names, location ranges, or sequence. Details of how to select appropriate genes and view gRNAs are included in **Workflow 3**. Once genes have been selected, gRNAs can be generated in one of two ways:

   (a) **View Targets**: In-depth analysis and user-specified export to CSV.
   (b) **Generate Library**: User-specified library settings with batch export to CSV.

In addition to the main gene targeting functionality, Main Window also contains smaller, useful features, including:

- **New Endonuclease**

- **NCBI Genome Browser**

- **NCBI Tool**

These features are discussed in greater detail below.

New Endonuclease: The New Endonuclease (Figure 5) feature allows the user to define a custom endonuclease, including its name, classification, PAM requirements, and gRNA sequence details (seed length, 5' length, 3' length, see **New Genome** section for definitions of these terms). Once submitted, the new endonuclease entry is saved into the CASPERinfo file that is located within the app directory. This tool is opened from the Main Window by clicking on the "Define New Endonuclease" button.

CASPER's on- and off-target algorithms can incorporate empirically-derived scoring matrices on an endonuclease-specific basis. For example, when scoring off-targets against spCas9 gRNAs, CASPER uses "HSU MATRIX-spCas9-2013" that is defined in CASPERinfo. New on- and off-target scoring matrices may be added to CASPERinfo manually, as long as they are formatted like the other datasets. When uploading a new endonuclease, the user is able to select on- and off-scoring matrices from dropdown menus automatically populated from the CASPERinfo file. This allows CASPER to retain its high-level predictive capabilities for on- and off-target scoring as new and improved datasets become available.

The user simply fills out the form, clicks "OK," and the new endonuclease definition is added to the CASPERinfo file. The user may then use their custom endonuclease in the creation of new CSPR files and subsequent analyses. This feature allows CASPER to nimbly adapt to emerging information as novel and engineered Cas enzymes are made available.

For a detailed demonstration of this process, please refer to **Workflow 4**.

Figure 5: New Endonuclease window.

NCBI Genome Browser: The Genome Browser (Figure 6) feature allows users to visualize the architecture and orientation of the genes they are targeting by use of NCBI's genome viewer. Simply select the desired annotation file from the dropdown menu in Main Window and navigate to the Genome Browser via the toolbar: Genome → Open Genome Browser.
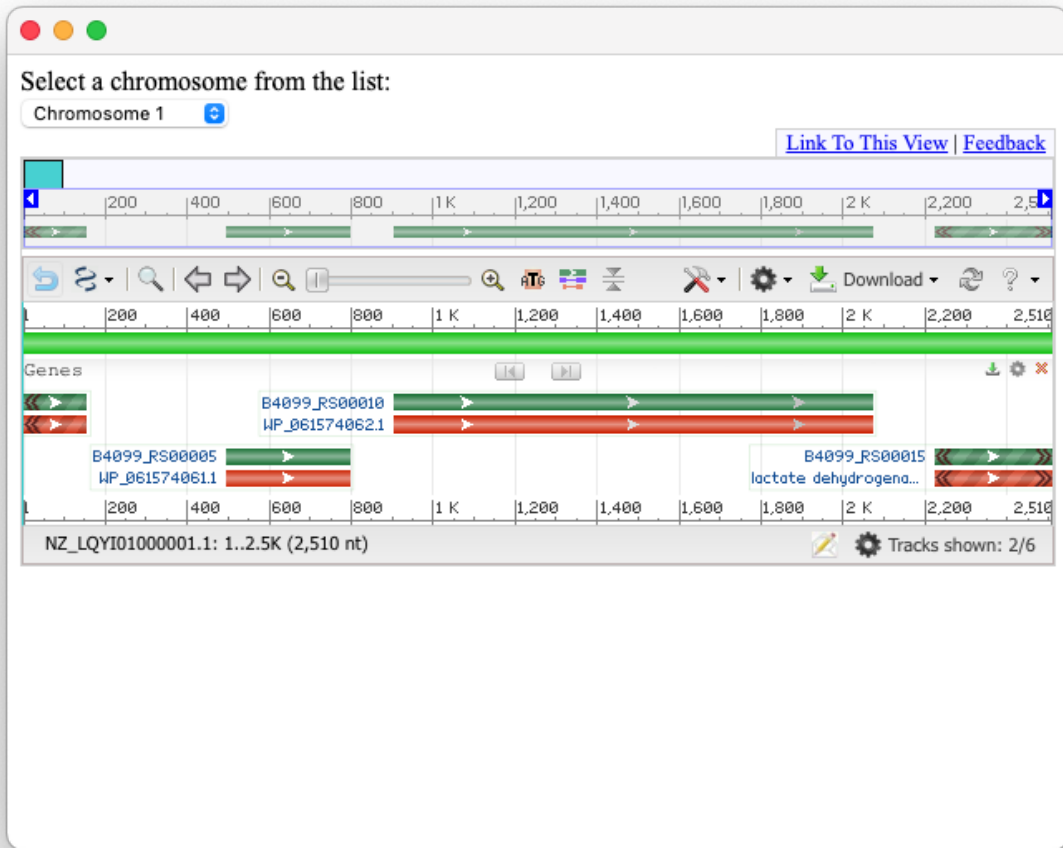
Figure 6: Genome Browser window.

NCBI Tool: The last and perhaps most useful side tool of the Main Window is the NCBI Tool, which is used for the automated retrieval of annotation and genomic sequence files from NCBI's databases (Figure 7). This allows users to query the vast NCBI GenBank and RefSeq collections without ever having to leave CASPER. A detailed demonstration for navigating this tool is included in **Workflow 2**.

It is *highly* recommended that users download GBFF files from the RefSeq collection if available, as these files tend to be better curated and contain more information. The gene feature and sequence search capabilities of CASPER cannot be guaranteed with annotation files from the GenBank collection.
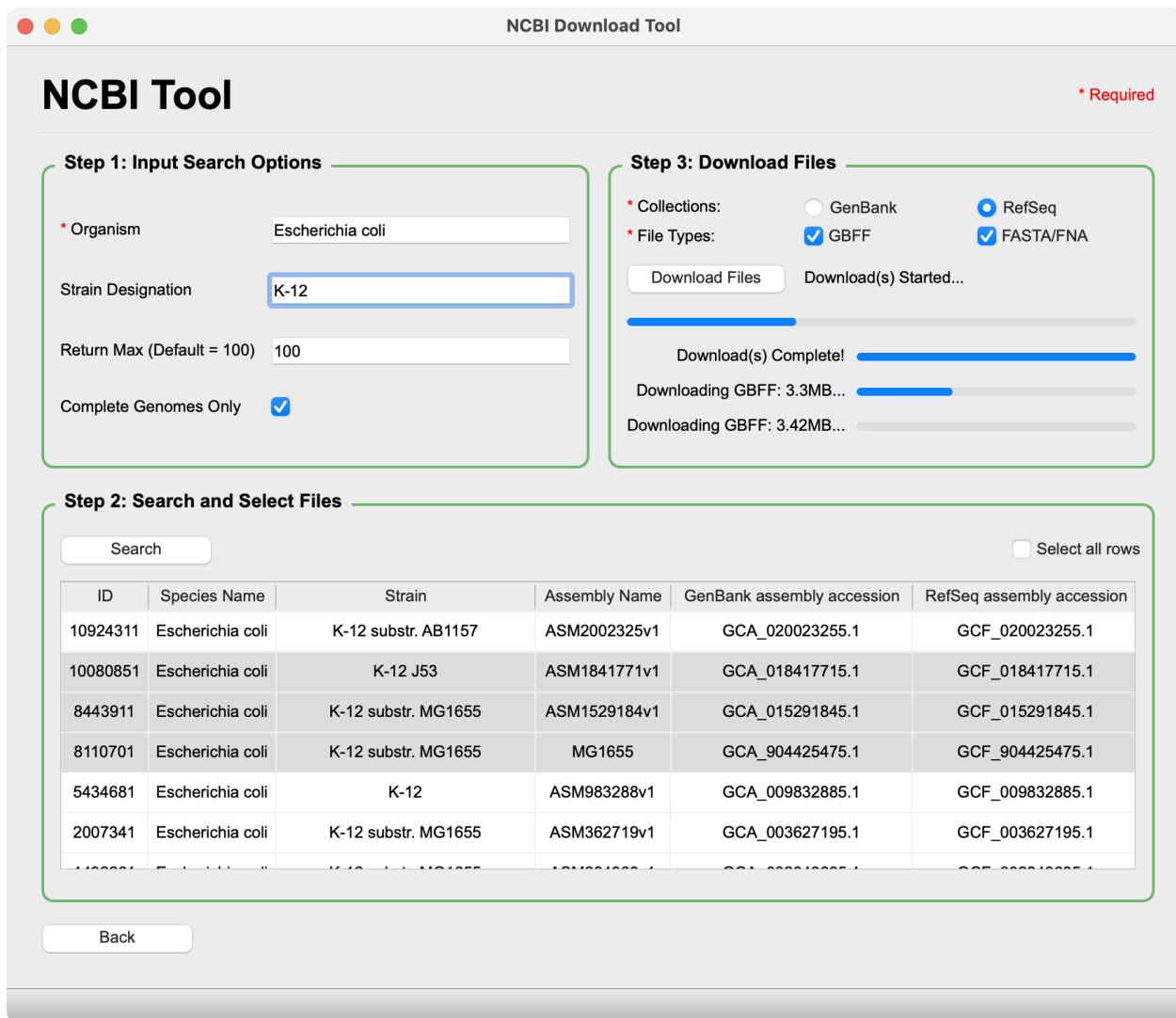
Figure 7: NCBI Tool window.

View Targets:

The View Targets portion of CASPER (Figure 8) is reached once a gene, location, or sequence search is performed and the "View Targets" button is clicked. Within this window, users are able to see all of the potential gRNAs for the selected gene or locus. In addition to multiple filtering options located directly above the gRNA table, there are several features within View Targets to aid in the assessment of highly active and specific guides:

- Gene Viewer

- Co-Targeting

- Export to CSV

- Off-Target Analysis
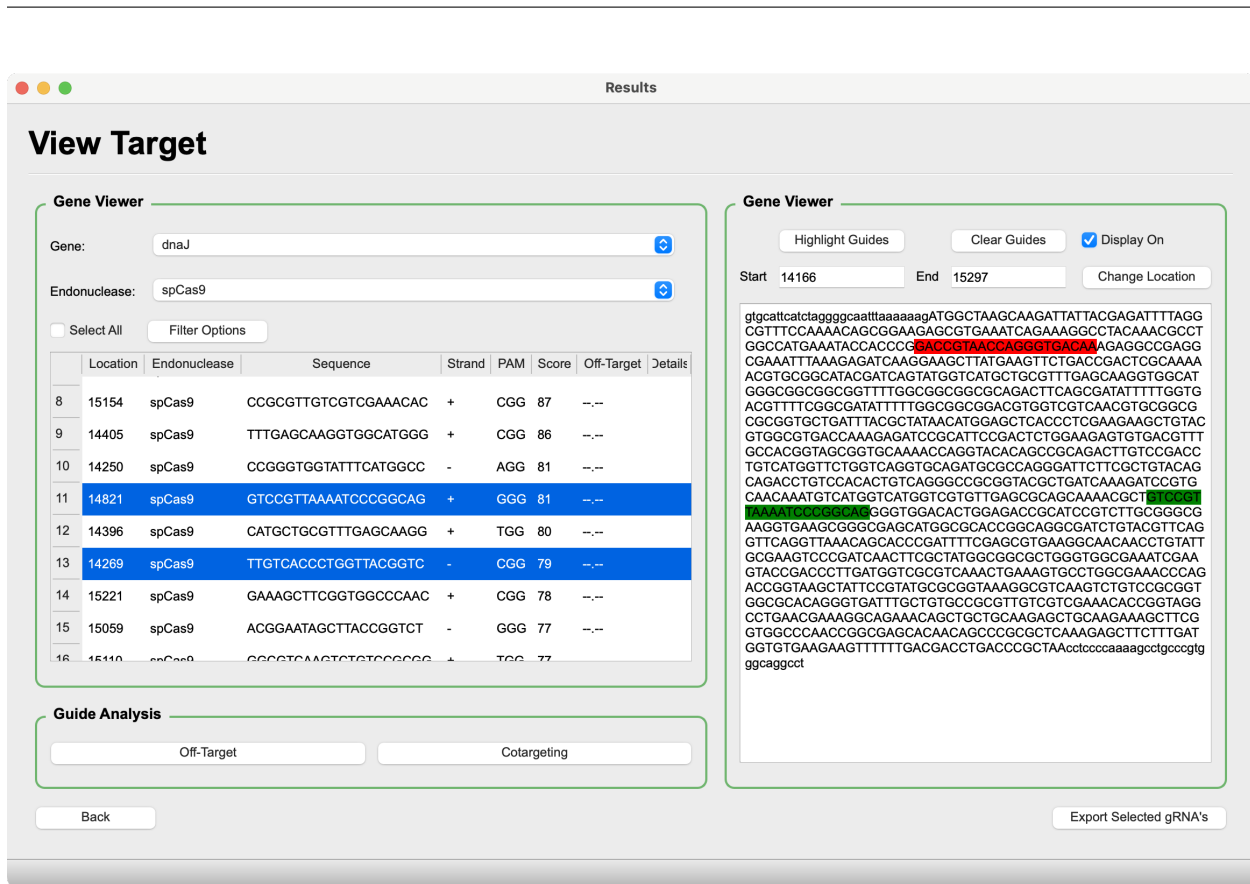
Figure 8: View Targets window.

Gene Viewer: Gene Viewer allows users to easily visualize their gRNAs from within the View Targets window. By default, the Gene Viewer is hidden to improve performance. To turn on Gene Viewer, simply click the "Display On" checkbox and the window will be populated with the appropriate gene sequence, along with 30 nt of padding around the gene (lowercase characters) to allow highlighting of gRNAs that appear at the gene's extrema.

Note: the Gene Viewer feature is populated based on the currently selected annotation file. If the annotation file selected does not match the organism selected, the Gene Viewer will display the wrong sequence. Also, if an annotation file isn't provided or is the wrong file type, the Gene Viewer will not work.

Co-Targeting: Certain Cas enzymes have similar enough PAMs that a single gRNA may be designed to simultaneously have activity with two or more Cas proteins at once, greatly reducing cloning efforts and providing more straightforward comparison between endonucleases. This is called Co-Targeting.

For example, because spCas9 and saCas9 have the 3' PAMs NGG and NNGRTT, respectively, genomic sequences adjacent to NGGRTT will be able to satisfy the PAM requirements of both endonucleases. These sequences can be said to be co-targets.

To perform Co-Targeting, users must first create CSPR files for the same organism using

all of the (compatible) endonucleases they wish to co-target with. Make sure all of the information entered into New Genome is the same for each organism, except for the endonuclease, so that CASPER will recognize the organism as available for Co-Targeting. Then, simply target a gene as normal and click on the "Co-Targeting" button in the View Targets window. CASPER will populate a table with all of the available Cas enzymes for that organism that are amenable to Co-Targeting. Lastly, highlight the desired enzymes and click "Submit" and guide RNAs will be populated in the View Targets table for use with each endonuclease as well as multiple endonucleases. To only view gRNAs that are compatible with multiple endonucleases, click "Filter Options" and check the "Co-Targets Only" box. For a detailed demonstration of Co-Targeting, please refer to **Workflow 5**.

Export to CSV: Export to CSV allows users to easily export gRNAs to a CSV file from the gRNA table in View Targets, as well as tables in **Multitargeting** and **Population Analysis**. To export, the user simply highlights the rows of the table they wish to export, specifies the destination directory and filename, and clicks "Submit". Optionally, users may choose to add a 3' or 5' motif to each of the gRNAs, such as a tRNA sequence to facilitate multiplexing. CSV files may easily be ported into a sheet-based editor like Excel or Google Sheets to facilitate facile manipulation or ordering of gRNAs.

Off-Target Analysis: Off-Target Analysis starts by highlighting the rows of the guides that the user wishes to analyze and then is launched by clicking the "Off-Target" button. A small window will pop up providing the user with various options for performing off-target analysis (Figure 9). Our proprietary Off-Target Algorithm is described in detail in our previous publication.

Once all of the fields have been filled out, click "Run", wait till the progress bar shows that the scoring has been completed, and click "Submit" to return to View Targets, where the off-target scores will be shown in the table.
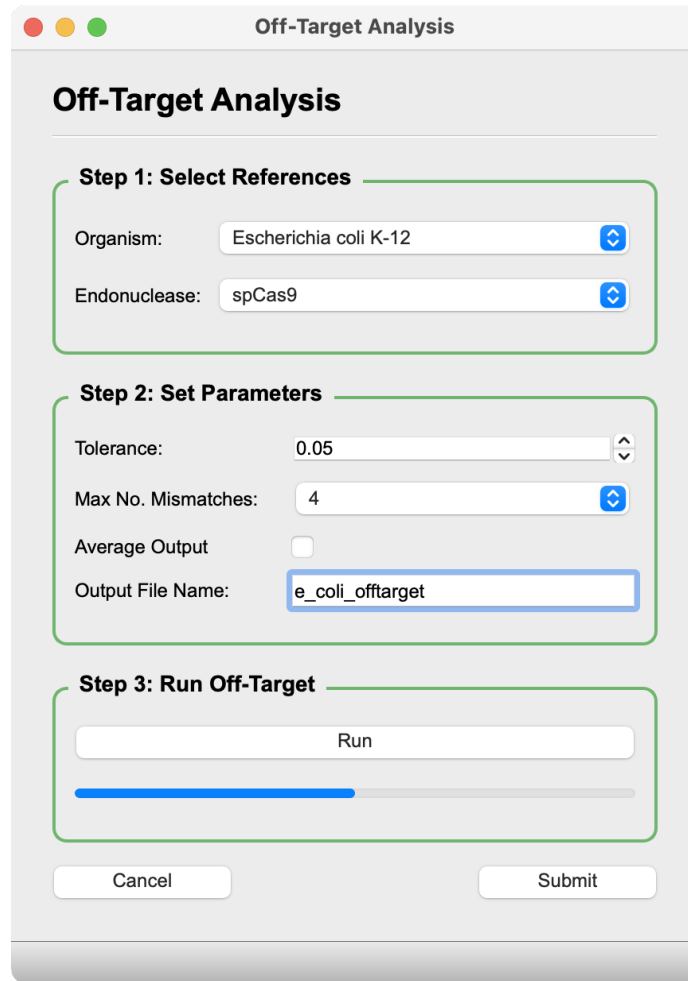
Figure 9: Off-Target Analysis window.

Terminology:

- <u>Organism</u>: Reference to be examined for off-target hits.

- <u>Endonuclease</u>: Endonuclease whose PAM will be used to locate off-target hits. (Ex. spCas9 = NGG)

- <u>Tolerance</u>: Minimum off-target score for an off-target hit (lower means less chance of activity). Sites with off-target scores less than this value will not be categorized as off-target hits, meaning they will be treated as 0.
  Note: computational burden increases as this value is decreased.

- <u>Max No. Mismatches</u>: Maximum number of mismatches relative to the gRNA within a putative off-target site that will be considered. All off-target sites with more mismatches than this value will not be considered for off-target analysis.
  Note: computational burden increases as this value is increased.

- Average Output: Outputs averaged off-target scores only. Default output format is detailed and gives an average score, along with a breakdown of each contributing putative off-target site.

- Check Non-Canonical PAMs: Allows users to search for off-target hits at sites with different PAMs than the canonical PAM(s) for the selected endonuclease.

Generate Library:

In Generate Library (Figure 10), users can quickly produce bespoke libraries of guides that adhere to specified criteria for a given gene or selection of genes. The criteria include library size and location, quality filtering options, and targeting restrictions. To run Generate Library, the user must first search for and select one or several genes in Step 3 of the **Main Window**. After submitting the selection, users may then click on "Generate Library" to continue to the Generate Library window. From there, the user simply fills out the fields with their desired parameters and clicks "Submit." A detailed video of this process can be found in **Workflow 6**.

Terminology:

- Number of Guides per Gene: Number of gRNAs that will attempt to be returned for each of the selected gRNAs. (Ex. if this option is set to 2 and 2 genes were selected, the library will contain 4 gRNAs, 2 for each gene.)
  Note: if the "Modify Parameters If No Target Found" option is not checked, CASPER may return less than the requested number of guides due to there not being enough that fit the specified conditions.

- Targeting Range: Percentage of the gene from which guides will attempt to be pulled. (Ex. if wanting only gRNAs from the back half of the gene, the user would target from 50 to 100.)

- Find Off-Targets: Check this to filter library guides based on off-target score.

- Minimum On-Target Score: Minimum on-target score for guides included in library. (Higher is better. Higher on-target score means higher predicted on-target activity)

- Maximum Off-Target Score: Maximum off-target score for guides included in library. (Lower is better. higher off-target score means higher predicted off-target activity)

- Space Between Guides: Minimum amount of space that must exist between guides included in library.

- 5' End Specificity: Specifies the 5' characters of the guides included in library. For example, to optimize gRNA expression under some promoters, 5' G residues are needed. Typing "G" in the box returns only guides starting with G. Input must be "ACGT" alphabet.

- Modify Parameters If No Target Found: Checking this box will relax the specified library guidelines in order to always fill the Number of Guides per gene requested.

Figure 10: Generate Library window.

Multitargeting:

Users may navigate to Multitargeting by clicking the "Multitargeting Analysis" button in the CASPER Navigation pane of the **Main Window**. Multitargeting (Figure 11) allows users to view repeated gRNA motifs across an entire genome of interest at both global and granular levels. This is accomplished by searching for gRNA seeds, which are the portions of gRNAs most imperative for target acquisition and cleavage. In spCas9, the seed is the first 16 nt after the PAM, leaving a variable "tail" region of 4 nt for a 20 nt gRNA. Searching for seeds effectively expands the search space for multitargeting and population analyses.

On the left side of the interface, users are able to interact with specific gRNA seeds by clicking and highlighting rows of the table. Figures are automatically populated for the seed corresponding to the selected row, revealing 1.) The number of repeats per chromosome/scaffold, and 2.) The location of the repeats within the selected genome. In the bottom right corner of the Multitargeting window, genome-scale statistics and figures are included to give users a high-level overview of the repeat landscape in the selected genome. A breakdown of all the figures is included below:

- Seed Analysis:

Note: If multiple rows are selected, only data from the last row selected will be used to populate these figures.

– Scaffold/Chromosome Viewer: Red lines show the location of repeats for the selected seed. The black boxes represent either chromosomes, scaffolds, or contigs, depending on how the FASTA file was formatted when the CSPR files were created. Detailed information about each repeat is shown above the viewer upon hover.

– Seed Distribution: Shows how many repeats are on each chromosome/scaffold/contig for the selected seed. If the number of chromosomes/scaffolds/contigs exceeds 10, some axis labels are removed to avoid labels overlapping, but the tick marks remain consistent with chromosome number.

• Global Analysis:

– Statistics Overview: Provides basic, genome-wide multitargeting statistics for the selected organism.

– Repeats per ID Number: Shows the total number of repeats for every seed in the selected genome. Seeds are arbitrarily assigned IDs.

– Seeds per Number Repeats: Shows the number of unique seeds per repeat number. Y axis is scaled logarithmically for readability.

• Seed Search Table: Summary of gRNA seeds that are present in the selected organism. To further analyze a given seed, simply highlight the appropriate row in the seed table and the appropriate figures are populated.

Terminology:

– Total Repeats: Total number of times the seed sequence is repeated across all organisms analyzed

– Avg. Repeats/Scaffold: Average number of times the seed sequence is repeated per scaffold/chromosome/contig (varies depending on how the organism's FASTA file is arranged)

– Consensus Sequence: Full length guide RNA sequence that appears most commonly amongst all occurrences of the given seed.

– % Consensus: Percentage of all seed repeats that have the tail of the consensus sequence.

For a detailed demonstration of how to perform and interpret multitargeting analyses, please refer to **Workflow 7**.
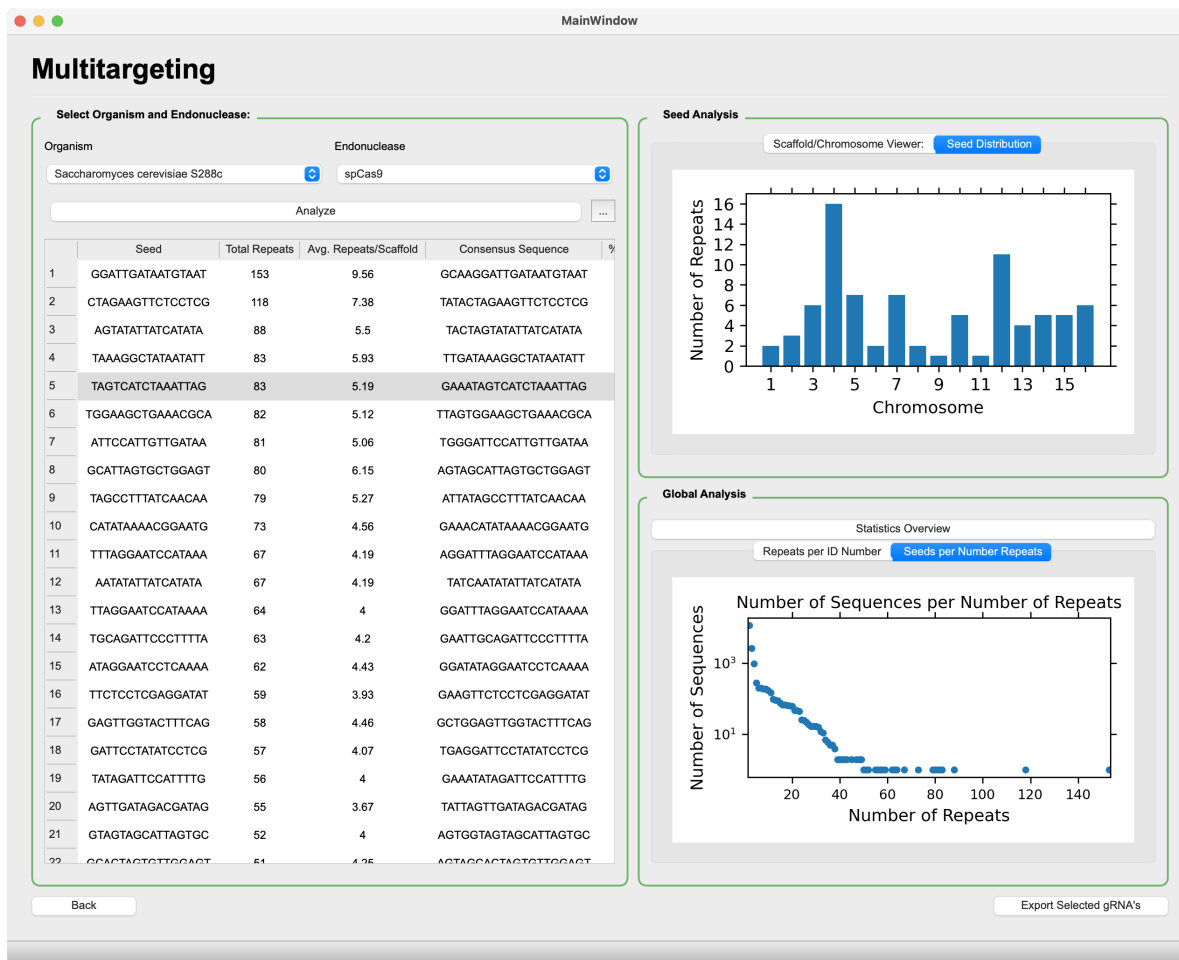
Figure 11: Multitargeting window.

Population Analysis:

Users may navigate to Population Analysis by clicking the "Population Analysis" button in the CASPER Navigation pane of the **Main Window**. Population analysis (Figure 12) allows users to select multiple organisms that have had CSPR files created for them and analyze the repeated gRNA seeds across the community. In this analysis, users will find:

- Shared Seed Heatmap: Pairwise distribution of repeated gRNA seeds for the selected organisms and endonuclease. Diagonals do not match color scale due to the high number of self-contained repeats relative to shared repeats, which skews the color scale. Hovering the cursor over tiles will show data callouts.

  Note: Axis indices correspond to the index of the selection table in the top left of the window.

- Seed Search Table: summary of gRNA seeds that are present in all of the organisms analyzed. To investigate the individual occurrences of a seed sequence, simply highlight the appropriate row in the seed table and click "Find Locations."

Terminology:

- – <u>% Coverage</u>: Percentage of analyzed organisms covered by the given seed sequence. (Ex. a coverage of 75% for an analysis of 4 organisms means that the seed appears in 3/4 of the organisms)

- – <u>Total Repeats</u>: Total number of times the seed sequence is repeated across all organisms analyzed

- – <u>Avg. Repeats/Scaffold</u>: Average number of times the seed sequence is repeated per scaffold/chromosome/contig (varies depending on how the organism's FASTA file is arranged)

- – <u>Consensus Sequence</u>: Full length guide RNA sequence that appears most commonly amongst all occurrences of the given seed.

- – <u>% Consensus</u>: Percentage of all seed repeats that have the tail of the consensus sequence.

To perform population analysis, select 1 or more CSPR databases from the table that are automatically populated from genomes previously analyzed in **New Genome**. Then, simply press "Analyze Organism(s)" to perform Population Analysis.

Note: Organisms must be analyzed with the same endonuclease within **New Genome** in order to be included in the same population analysis.

For a detailed demonstration of how to perform and interpret population analyses, please refer to **Workflow 7**.
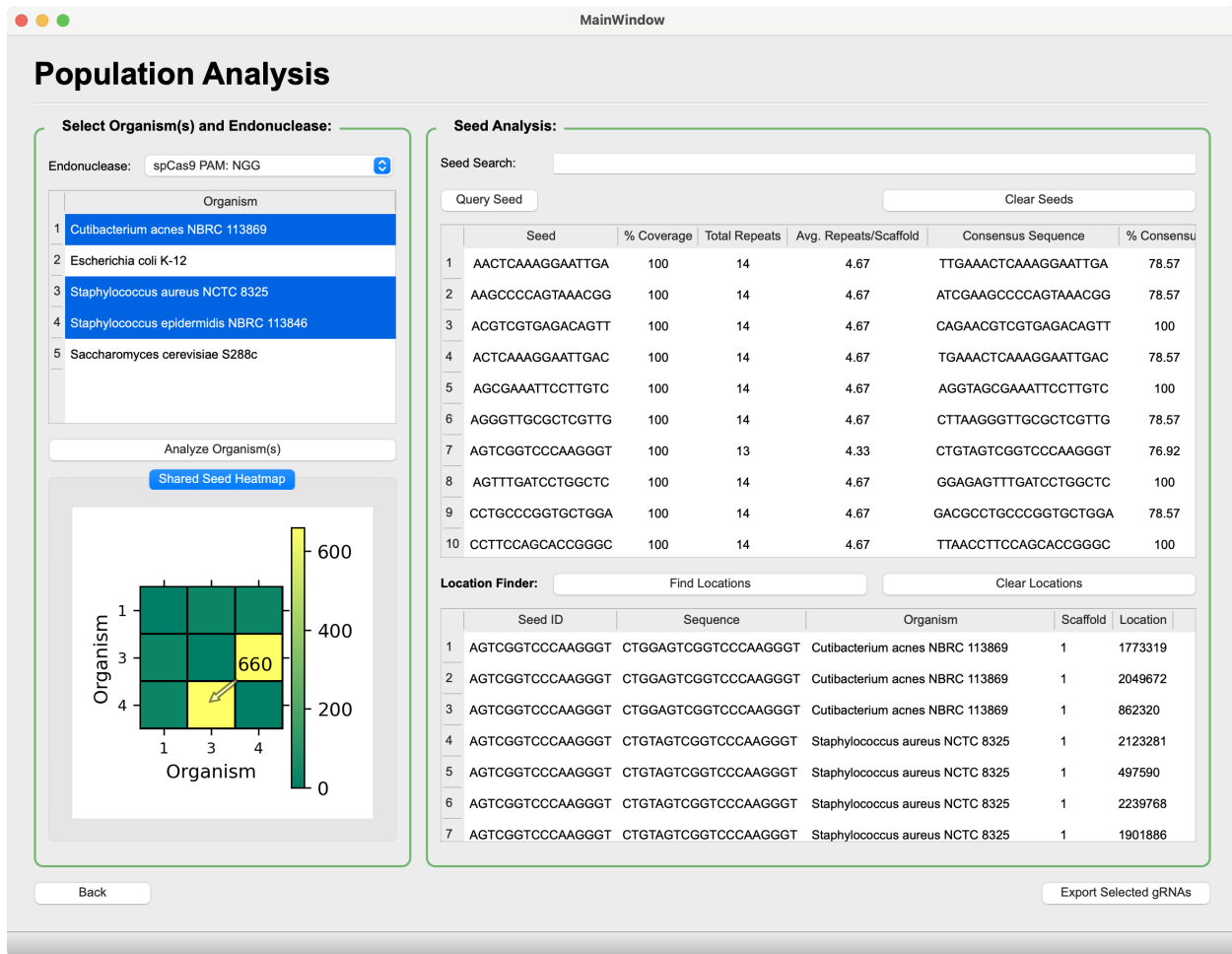
Figure 12: Population Analysis window.

YOUTUBE TUTORIALS

[Workflow 1: Installing CASPER](#)

[Workflow 2: Analyze a New Genome](#)

[Workflow 3: Perform Single Gene/Locus gRNA Generation](#)

[Workflow 4: Define New Endonuclease](#)

[Workflow 5: Perform Co-Targeting Analysis](#)

[Workflow 6: Perform gRNA Library Generation](#)

[Workflow 7: Perform Multitargeting/Population Analyses](#)