ACS Synthetic Biology

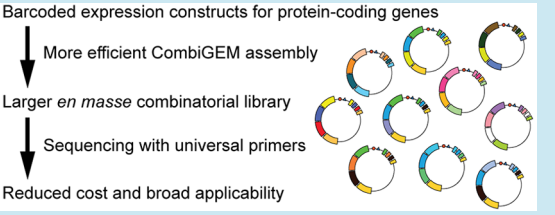
pubs.acs.org/synthbio
Technical Note

Improved Combinatorial Assembly and Barcode Sequencing for Gene-Sized DNA Constructs

Diana Hernandez Hernandez, Lin Ding, Ayako Murao, Lukas R. Dahlin, Gabriella Li, Kathleen L. Arnolds, Melissa Amezola, Amit Klein, Aishwarya Mitra, Sonia Mecacci, Jeffrey G. Linger, Michael T. Guarnieri, and Yo Suzuki*

Cite This: ACS Synth. Biol. 2023, 12, 2778–2782

ABSTRACT: Synergistic and supportive interactions among genes can be incorporated in engineering biology to enhance and stabilize the performance of biological systems, but combinatorial numerical explosion challenges the analysis of multigene interactions. The incorporation of DNA barcodes to mark genes coupled with next-generation sequencing offers a solution to this challenge. We describe improvements for a key method in this space, CombiGEM, to broaden its application to assembling typical gene-sized DNA fragments and to reduce the cost of sequencing for prevalent small-scale projects. The expanded reach of the method beyond currently targeted small RNA genes promotes the discovery and incorporation of gene synergy in natural and engineered processes such as biocontainment, the production of desired compounds, and previously uncharacterized fundamental biological mechanisms.



KEYWORDS: *combinatorial genetics en masse, enzymatic ligation assisted by nucleases, next-generation sequencing, genetic interaction, multigene synergy for biocontainment, epistasis*

INTRODUCTION

Synthetic biology provides powerful tools to transform biological discoveries and applications, but the lack of understanding of biological systems is often an impediment for the successful implementation of the tools. A major source of uncertainty is multigene interactions.[1] Engineered constructs and genomic edits can interact among themselves or with complex cellular machineries in unpredictable ways. In organism engineering, exogenous genes, including artificially designed genes, can create novel gene combinations to potentiate the organism's useful features for target applications. For example, a promising approach is to combine multiple engineered constructs to establish robust biocontainment systems for large culture scales in industrial settings.[2] However, determining the combinations of DNA constructs to maximize the effect is challenging.

To facilitate the analysis of combinatorial constructs, a powerful method termed combinatorial genetics *en masse* (CombiGEM)[3] was developed. In this method, the recursive cloning of DNA-barcoded modules via specific designated restriction sites within the modules (between the construct and the DNA barcode in each module) results in the accumulation of the DNA constructs on one side and the DNA barcodes on the other side of the restriction sites (Figure 1a). This recursive process is conducted *en masse* with a pool of insert fragments in each round, and the resulting multimodule constructs are transferred to an organism or system of interest for phenotype characterization. In the experiment, the concatenated barcodes analyzed with next-generation sequencing (NGS) identify the multiple modules, for example, in each cell in the mixed population of transformed cells in a target organism in tested conditions. This method has been shown to be highly effective for cloning and analyzing multiple microRNA constructs and guide RNA constructs for CRISPR.[4,5] However, unlike in methods involving gene shuffling within organisms,[6] the inefficiency in cloning larger constructs hindered the generation of diverse combinatorial genotypes in CombiGEM in our hand (Table S1). This Technical Note describes a set of improvements to make CombiGEM more efficient for fragment sizes typical for protein-coding genes, with a range from 463 to 1538 bp examined, and make it cost-effective in common scenarios where a small number of *en masse* populations are analyzed.

RESULTS AND DISCUSSION

Incorporating a Single-Copy Vector to Accommodate Deleterious Sequences. In a workflow to generate DNA constructs *en masse* in an *Escherichia coli* population and then

Received: March 28, 2023
Published: August 15, 2023

https://doi.org/10.1021/acssynbio.3c00183
ACS Synth. Biol. 2023, 12, 2778–2782

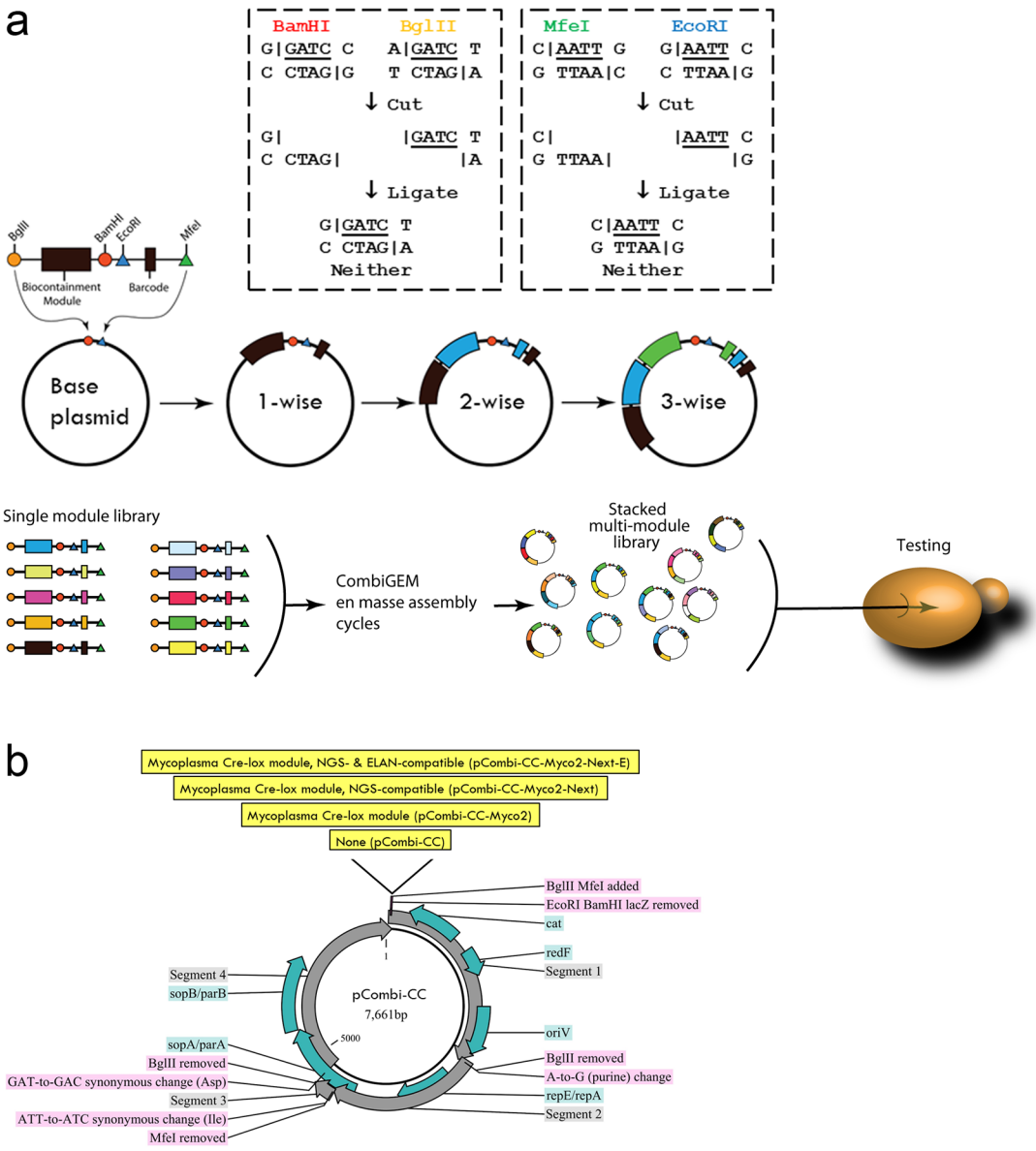


Figure 1. Improvements of the CombiGEM process. (a) CombiGEM procedure. The CombiGEM method uses the *Bam*HI and *Eco*RI sites on the vector and the compatible *Bgl*II and *Mfe*I sites in the insert fragments for cloning. The ligated sites are not recognized by any of the enzymes. The inserts bring new *Bam*HI and *Eco*RI sites within to enable the recursive assembly of modules. The process is run *en masse* with a pool of inserts. The resulting mixture of multimodule constructs are used in experiments. (b) pCombi-CC vectors. The *Bam*HI, *Bgl*II, *Eco*RI, and *Mfe*I restriction sites were engineered in the expandable single-copy vector pCC1BAC (Lucigen). The added *Bgl*II and *Mfe*I sites accept incoming fragments. To introduce functionality such as integration of the construct into the genome of *Mycoplasma mycoides* JCVI-syn3.0,[9] genetic tools are introduced into these sites along with the first barcode (position 0) to identify the tools (such as the set for mycoplasma Cre-lox integration). To make the vector compatible with the ELAN process (pCombi-CC-Myco2-Next-E), the *Bgl*II and *Mfe*I sites were disrupted. The gene *cat* encodes chloramphenicol acetyltransferase. *oriV* enables a high-copy replication when *trfA* gene product is provided in the cell (Lucigen). *repE/repA*, *sopA/parA*, *sopB/parB*, and *ori2* (not shown) participate in the replication and partition of F-factor replicon.

transfer the constructs into a target organism of interest, the diversity of DNA constructs must be retained during the *E. coli* phase, despite any toxicity. An approach to alleviate the toxicity of any DNA construct is to reduce the copy number of the construct, achievable with a single-copy vector such as pCC1BAC (Lucigen). The CombiGEM method requires four restriction sites in specific places, for example, the *Bam*HI and *Eco*RI sites on the vector and the compatible *Bgl*II and *Mfe*I sites, respectively, in the insert fragments for cloning (Figure 1a). To enable the process with pCC1BAC, we removed five unwanted restriction sites from the vector and added two sites (*Bgl*II and *Mfe*I; Supplementary Methods; Figure 1b). The "domesticated" base plasmid pCombi-CC (copy control) was then used to demonstrate *en masse* multigene construction in *E. coli* (Figure S1). This vector can be induced to increase its copy number (Lucigen) and make abundant DNA to facilitate the CombiGEM process.

Enhancing Compatibility with NGS. To genotype concatenated barcodes, amplicon sequencing is used.[7] This approach typically generates flanking sequences that are different from those in the standard Illumina TruSeq process; therefore, project-specific primers and dedicated flow cells are needed. To reduce costs for small experiments, we developed pCombi-CC-Next vectors and an associated process integrat-

https://doi.org/10.1021/acssynbio.3c00183
ACS Synth. Biol. 2023, 12, 2778−2782

ing amplicon sequencing with standard Illumina primers. These vectors contain part of the TruSeq Index Adapter sequence. The incoming gene fragments contain part of the TruSeq Universal Adapter sequence. Hence, a quick single PCR reaction with short primers can complete the construction of libraries to sequence with universal primers. These primers also introduce indexes for multiplexing to mark each mixture of constructs generated *en masse* (Figure S2). When libraries are compatible with standard primers for Illumina processes, they can be included in a run with other samples to achieve a marked reduction in cost. Also, they can be brought to any facility with sequencing platforms where standard Illumina primers are used for quick turnaround. Using pCombi-CC-Next vectors, we conducted CombiGEM processes with a pool of four mock biocontainment modules (also adjusted for NGS) of the sizes 463, 542, 867, and 1538 bp and demonstrated genotyping for DNA barcodes with Illumina NovaSeq sequencing (Table S2; Table S3). The sequencing cost was $7 USD for approximately 1 million reads representing each library. The small cost enables using CombiGEM in an education setting, as part of a common laboratory analysis, or in genome-scale studies.

Increasing Cloning Efficiency. When combinatorial DNA constructs are established in *E. coli* before introduction into the final target organism, inefficient *E. coli* transformation can be a major bottleneck, especially with larger inserts. Following a standard ligation protocol, we obtained roughly 6000 colonies per μg of vector DNA (30−150 colonies for 5−25 ng typically used per plate). This number is barely sufficient for low-level combinatorics with a few variable inserts. To increase transformation efficiency, we implemented the enzymatic ligation assisted by nucleases (ELAN) method.[8] Here, restriction enzymes that digest the ends of the fragments-to-be-ligated are introduced into the ligation reaction to prevent the circularization of the inserts and make more of them available for ligation with the vector. This strategy is effective when the vector and the insert are combined with compatible but different enzyme sites such as *Bam*HI and *Bgl*II, as in the CombiGEM method.

Using this method, we obtained more colonies (data not shown), but the ELAN method created a new problem. Residual *E. coli* genomic DNA (gDNA) contained in plasmid preparations for the vector side (Table S4) resulted in the cloning of genomic fragments, as was evident in the vector-only control samples for the CombiGEM assembly process. Gel purification performed for inserts was effective for removing gDNA, but it was not suitable for the vector side that could be an *en masse* generated pool of different sized samples.

The vector was digested with *Bam*HI and *Eco*RI, but the ends of the cloned gDNA were always *Bgl*II and *Mfe*I sites that were activated during the ELAN process with *Bgl*II and *Mfe*I. The likely reason for not cloning gDNA fragments ending with *Bam*HI and *Eco*RI sites was that these sites were inactivated during the phosphatase treatment to block the self-ligation of the vector. This meant that dephosphorylating *Bgl*II and *Mfe*I sites also should be effective for preventing gDNA cloning. When the vector preparation was digested with four enzymes (*Bam*HI, *Bgl*II, *Eco*RI, and *Mfe*I) rather than only two (*Bam*HI and *Eco*RI), followed by phosphatase treatment before introducing the vector into the ELAN process, we did not detect the cloning of gDNA fragments in the 1-wise cycle. Simultaneously, we achieved an improvement in transformation efficiency by 8-fold in 1-wise assembly, also producing high-level counts in later assemblies (Table S5). However, gDNA contamination was detected in cycles 2 and 3 and became prevalent in cycle 4 (Table S6). Therefore, we performed two rounds of 4-enzyme digestion and phosphatase treatment. The doubly digested pCombi-CC-Next vector produced a ∼100-fold lower background (colonies ascribed to only the vector material) than that singly digested in *E. coli* transformation (Table S7). The "singly digested" sample showed gDNA cloning in the second round of a modified CombiGEM process with no inserts added, whereas the doubly digested sample did not (Table S8).

Exonuclease V (RecBCD) digests linear but not circular DNA. It is marketed as a tool to remove gDNA from plasmid preparations (New England Biolabs), as gDNA is often fragmented or partially synthesized. To evaluate the efficacy in a single-round, reproducible CombiGEM experiment, we spiked in purified *E. coli* gDNA into our vector sample. The treatment with exonuclease V reduced the number of colonies produced after *E. coli* transformation (Table S9). When the plasmids were prepared and sequenced, we found gDNA inserted for 0% of the colonies in the exonuclease-treated set and 100% in the untreated set (n = 16; Table S10). Taken together, we have identified two approaches for reducing gDNA cloning. When the vector was allowed to close by itself, we occasionally obtained truncated vector clones (Table S8; Table S10). The reason may be star activity from restriction enzymes, incomplete plasmid replication, or running the ELAN process without inserts, creating free ends at unintended positions in the vector. The protocol is currently not optimized for reducing the number of instances of truncation.

To see if the newly developed tools work together, we conducted four cycles of the ELAN-enhanced CombiGEM process with the measures for preventing gDNA fragment cloning and analyzed the products using NGS. We observed the expected progressive clustering of barcodes in each cycle (Table S11; Figure S3). The combinatorial diversity also increased, but there were considerable biases in insert incorporation. Notably, the longest 1538-bp insert appeared to be underrepresented in most assemblies, and constructs carrying multiple copies of this insert were rare. However, there were sufficiently many constructs with one copy to enable studies not involving gene dosage or permutation. The result may be improved by adjusting the abundances of inserts to favor the incorporation of longer ones.

CONCLUSIONS

The improvements described enable a CombiGEM workflow in which thousands of *en masse*-assembled constructs with typically sized genes are generated and tracked with ease. This workflow suits prevalent scenarios in biological studies such as identifying 4-gene interactions with 10 possible genes (210 combinations). The larger community effort to understand complex gene interactions will propel synthetic biology in many systems where gene interactions play a role.

METHODS

Standard recombinant DNA methods were used. The details are found in the Supplementary Methods document.[10−13]

https://doi.org/10.1021/acssynbio.3c00183
ACS Synth. Biol. 2023, 12, 2778−2782

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.3c00183.

Supplementary Methods; Figure S1. Initial tests of the pCombi-CC vector in a CombiGEM procedure; Figure S2. Sequences around the cloning junctions in a pCombi-CC-Next clone; Figure S3. Graphical representation of genotypes produced in 4-cycle CombiGEM processes revealed by DNA barcode sequencing; Table S1. Insert size affecting CombiGEM efficiency; Table S2. Next-generation sequencing result for 2-wise assembly; Table S3. Raw counts from the analysis of a 2-wise assembly; Table S4. *E. coli* genomic DNA in plasmid preparation; Table S5. Colony counts with the ELAN process; Table S6. Cloning of gDNA fragments with the ELAN-integrated CombiGEM process; Table S7. Colony counts for singly and doubly digested vector preparation with gDNA contamination; Table S8. Cloning of gDNA fragments into singly and doubly digested vectors; Table S9. Exonuclease treatment to remove gDNA; Table S10. Reduction of gDNA cloning with exonuclease V treatment; Table S11. CombiGEM processes with gDNA reduction; Table S12. List of primers used (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Yo Suzuki − *Synthetic Biology and Bioenergy Group, J. Craig Venter Institute, La Jolla, California 92037, United States;* orcid.org/0000-0002-2797-6922; Email: ysuzuki@jcvi.org

Authors

Diana Hernandez Hernandez − *Synthetic Biology and Bioenergy Group, J. Craig Venter Institute, La Jolla, California 92037, United States;* orcid.org/0000-0002-3914-3581

Lin Ding − *Synthetic Biology and Bioenergy Group, J. Craig Venter Institute, La Jolla, California 92037, United States*

Ayako Murao − *Synthetic Biology and Bioenergy Group, J. Craig Venter Institute, La Jolla, California 92037, United States*

Lukas R. Dahlin − *National Renewable Energy Laboratory, Golden, Colorado 80401, United States*

Gabriella Li − *National Renewable Energy Laboratory, Golden, Colorado 80401, United States*

Kathleen L. Arnolds − *National Renewable Energy Laboratory, Golden, Colorado 80401, United States*

Melissa Amezola − *Synthetic Biology and Bioenergy Group, J. Craig Venter Institute, La Jolla, California 92037, United States*

Amit Klein − *Synthetic Biology and Bioenergy Group, J. Craig Venter Institute, La Jolla, California 92037, United States; Department of Bioengineering, University of California San Diego, La Jolla, California 92093, United States*

Aishwarya Mitra − *Synthetic Biology and Bioenergy Group, J. Craig Venter Institute, La Jolla, California 92037, United States; Department of Bioengineering, University of California San Diego, La Jolla, California 92093, United States*

Sonia Mecacci − *Synthetic Biology and Bioenergy Group, J. Craig Venter Institute, La Jolla, California 92037, United States*

Jeffrey G. Linger − *National Renewable Energy Laboratory, Golden, Colorado 80401, United States*

Michael T. Guarnieri − *National Renewable Energy Laboratory, Golden, Colorado 80401, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.3c00183

Author Contributions

D.H.H., L.D., and Y.S. designed and performed the experiments. A.M., L.R.D., G.L., K.L.A., M.A., A.K., A.M., S.M., J.G.L., and M.T.G. contributed technical assistance. D.H.H. and Y.S. drafted the manuscript, and all authors edited the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Nozomu Yachie for helpful discussion. This work was supported by U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomic Science Program under Secure Biosystems Design Science Focus Area IMAGINE BioSecurity: Integrative Modeling and Genome-scale Engineering for Biosystems Security, under contract number DE-AC36-08GO28308. This journal article was developed based upon funding from the Alliance for Sustainable Energy, LLC, Managing and Operating Contractor for the National Renewable Energy Laboratory for the U.S. Department of Energy. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

■ REFERENCES

(1) Suzuki, Y.; St Onge, R. P.; Mani, R.; King, O. D.; Heilbut, A.; Labunskyy, V. M.; Chen, W.; Pham, L.; Zhang, L. V.; Tong, A. H. Y.; Nislow, C.; Giaever, G.; Gladyshev, V. N.; Vidal, M.; Schow, P.; Lehár, J.; Roth, F. P. Knocking out Multigene Redundancies via Cycles of Sexual Assortment and Fluorescence Selection. *Nat. Methods* **2011**, *8* (2), 159−164.

(2) Arnolds, K. L.; Dahlin, L. R.; Ding, L.; Wu, C.; Yu, J.; Xiong, W.; Zuniga, C.; Suzuki, Y.; Zengler, K.; Linger, J. G.; Guarnieri, M. T. Biotechnology for Secure Biocontainment Designs in an Emerging Bioeconomy. *Curr. Opin. Biotechnol.* **2021**, *71*, 25−31.

(3) Cheng, A. A.; Ding, H.; Lu, T. K. Enhanced Killing of Antibiotic-Resistant Bacteria Enabled by Massively Parallel Combinatorial Genetics. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (34), 12462−12467.

(4) Wong, A. S. L.; Choi, G. C. G.; Cui, C. H.; Pregernig, G.; Milani, P.; Adam, M.; Perli, S. D.; Kazer, S. W.; Gaillard, A.; Hermann, M.; Shalek, A. K.; Fraenkel, E.; Lu, T. K. Multiplexed Barcoded CRISPR-Cas9 Screening Enabled by CombiGEM. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (9), 2544−2549.

(5) Wong, A. S. L.; Choi, G. C. G.; Cheng, A. A.; Purcell, O.; Lu, T. K. Massively Parallel High-Order Combinatorial Genetics in Human Cells. *Nat. Biotechnol.* **2015**, *33* (9), 952−961.

(6) Celaj, A.; Gebbia, M.; Musa, L.; Cote, A. G.; Snider, J.; Wong, V.; Ko, M.; Fong, T.; Bansal, P.; Mellor, J. C.; Seesankar, G.; Nguyen, M.; Zhou, S.; Wang, L.; Kishore, N.; Stagljar, I.; Suzuki, Y.; Yachie,

N.; Roth, F. P. Highly Combinatorial Genetic Interaction Analysis Reveals a Multi-Drug Transporter Influence Network. *Cell Syst.* **2020**, *10* (1), 25−38.e10.
(7) Caporaso, J. G.; Lauber, C. L.; Walters, W. A.; Berg-Lyons, D.; Lozupone, C. A.; Turnbaugh, P. J.; Fierer, N.; Knight, R. Global Patterns of 16S rRNA Diversity at a Depth of Millions of Sequences per Sample. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 4516−4522.
(8) Cost, G. J. Enzymatic Ligation Assisted by Nucleases: Simultaneous Ligation and Digestion Promote the Ordered Assembly of DNA. *Nat. Protoc* **2007**, *2* (9), 2198−2202.
(9) Hutchison, C. A.; Chuang, R.-Y.; Noskov, V. N.; Assad-Garcia, N.; Deerinck, T. J.; Ellisman, M. H.; Gill, J.; Kannan, K.; Karas, B. J.; Ma, L.; Pelletier, J. F.; Qi, Z.-Q.; Richter, R. A.; Strychalski, E. A.; Sun, L.; Suzuki, Y.; Tsvetanova, B.; Wise, K. S.; Smith, H. O.; Glass, J. I.; Merryman, C.; Gibson, D. G.; Venter, J. C. Design and Synthesis of a Minimal Bacterial Genome. *Science* **2016**, *351* (6280), No. aad6253.
(10) Kostylev, M.; Otwell, A. E.; Richardson, R. E.; Suzuki, Y. Cloning Should Be Simple: *Escherichia coli* DH5α-Mediated Assembly of Multiple DNA Fragments with Short End Homologies. *PloS one* **2015**, *10* (9), No. e0137466.
(11) Ding, L.; Brown, D. M.; Glass, J. I. Rescue of Infectious Sindbis Virus by Yeast Spheroplast-Mammalian Cell Fusion. *Viruses* **2021**, *13* (4), 603.
(12) Engler, C.; Gruetzner, R.; Kandzia, R.; Marillonnet, S. Golden Gate Shuffling: A One-Pot DNA Shuffling Method Based on Type IIs Restriction Enzymes. *PLoS One* **2009**, *4* (5), No. e5553.
(13) Zhou, P.; Chan, B. K. C.; Wan, Y. K.; Yuen, C. T. L.; Choi, G. C. G.; Li, X.; Tong, C. S. W.; Zhong, S. S. W.; Sun, J.; Bao, Y.; Mak, S. Y. L.; Chow, M. Z. Y.; Khaw, J. V.; Leung, S. Y.; Zheng, Z.; Cheung, L. W. T.; Tan, K.; Wong, K. H.; Chan, H. Y. E.; Wong, A. S. L. A Three-Way Combinatorial CRISPR Screen for Analyzing Interactions among Druggable Targets. *Cell Reports* **2020**, *32* (6), 108020.

https://doi.org/10.1021/acssynbio.3c00183
ACS Synth. Biol. 2023, 12, 2778−2782