SYNTHETIC BIOLOGY

# Synthetic sequence entanglement augments stability and containment of genetic information in cells

Tomasz Blazejewski[1,2]*, Hsing-I Ho[1]*, Harris H. Wang[1,3]†

**In synthetic biology, methods for stabilizing genetically engineered functions and confining recombinant DNA to intended hosts are necessary to cope with natural mutation accumulation and pervasive lateral gene flow. We present a generalizable strategy to preserve and constrain genetic information through the computational design of overlapping genes. Overlapping a sequence with an essential gene altered its fitness landscape and produced a constrained evolutionary path, even for synonymous mutations. Embedding a toxin gene in a gene of interest restricted its horizontal propagation. We further demonstrated a multiplex and scalable approach to build and test >7500 overlapping sequence designs, yielding functional yet highly divergent variants from natural homologs. This work enables deeper exploration of natural and engineered overlapping genes and facilitates enhanced genetic stability and biocontainment in emerging applications.**

Protein-encoding information is stored in DNA as a series of trinucleotide codons. Because protein translation can occur in one of six coding frames, multiple proteins could in principle be produced from different frames of a single DNA sequence. Empirically, such overlapping genes are widely found in biology from bacteria to humans (*1*–*4*). Across microbial genomes, overlapping genes are estimated to make up almost one-third of all coding sequences (*5*). Although partial overlaps are more typical, many completely overlapped genes have been described (*1*, *6*) including an example where three different proteins are translated from separate frames of the same mRNA (*7*).

An important consequence of overlapping genes is that mutations will affect all protein products simultaneously. Mutations that would otherwise be neutral in one frame may no longer be permitted if they create deleterious mutations in another frame, which constitutes a mechanism to preserve sequence fidelity (*8*). In fact, biological systems experiencing very high mutation rates, such as viruses, tend to more frequently contain overlapping genes (*9*). Past synthetic efforts to maintain DNA sequence fidelity have focused on reducing background mutation rates (*10*, *11*), eliminating mutation-prone sequences (*12*), or increasing mutation surveillance and correction (*13*). To prevent escape of recombinant DNA into the wild, various biocontainment strategies have also been developed (*14*).

Inspired by naturally overlapping genes that safeguard against mutations, we devised a platform, Constraining Adaptive Mutations using Engineered Overlapping Sequences (CAMEOS), to computationally design and experimentally test de novo overlapping genes (Fig. 1A). The overall computational objective is to identify co-encoding variants of two proteins of interest that share the same DNA sequence while minimizing disruptive residue changes in each protein sequence. Protein function can be disrupted by individual residue substitutions, insertions, or deletions, as well as by changes in long-range interactions between residue pairs. The CAMEOS algorithm addresses both considerations in two steps (Fig. 1B and fig. S1) (*15*). Briefly, a dynamic programming algorithm first generates a double-encoding solution that is optimal according to a hidden Markov model (HMM). High-performing suboptimal solutions are subsequently generated by a stochastic backtrace procedure. These HMM-derived solutions are used as seeds to the second step, in which pairwise long-range residue interactions

[1]Department of Systems Biology, Columbia University, New York, NY, USA. [2]Integrated Program in Cellular, Molecular, and Biomedical Studies, Columbia University, New York, NY, USA. [3]Department of Pathology and Cell Biology, Columbia University, New York, NY, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: hw2429@columbia.edu

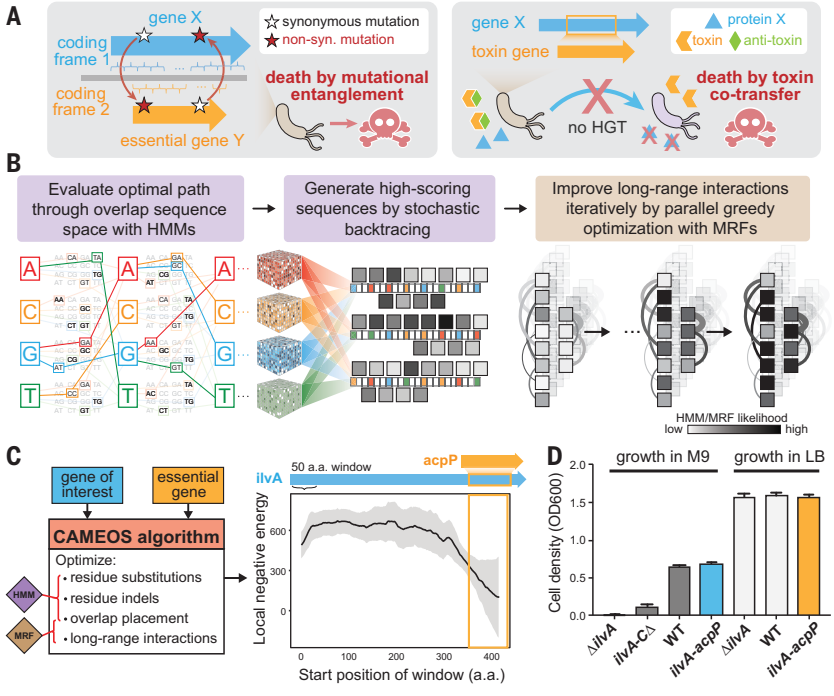


**Fig. 1. CAMEOS platform for designing overlapping genes.** (**A**) Schematic of mutational restriction or horizontal transfer confinement due to sequence entanglement of two genes. (**B**) The CAMEOS algorithm constructs a high-dimensional tensor (colored cubes) parameterizing the cost of paths (colored lines) through sequence space. These paths are then sampled probabilistically through a stochastic backtrace to form a population of sequences whose long-range interactions (gray arcs) are then optimized greedily and iteratively. (**C**) Parameters optimized by CAMEOS, with a schematic of co-encoded *ilvA* and *acpP* genes and the local negative energy of IlvA shown at the right. (**D**) Growth of a genomically encoded *ilvA-acpP* variant (IA-1) compared to that of control strains and wild-type (WT) cells after 14 hours in M9 minimal and LB rich media. The *ilvA*-CΔ strain is a Δ*ilvA* derivative with a plasmid expressing a C-terminal *ilvA* truncation variant at the overlap (residue 347). Data are means ± SEM from three independent biological replicates.

modeled by a Markov random field (MRF) (*16*, *17*) are iteratively optimized (fig. S2). We find that minimizing disruptions to long-range interactions is key for generating functional proteins (fig. S3 and table S1). Finally, synonymous mutations are made upstream of the overlap DNA region to optimize the Shine-Dalgarno sequence (*18*) for improved translation of the embedded gene.

We evaluated CAMEOS by designing and testing synthetic overlaps of essential and biosynthetic genes (Fig. 1C). Because certain protein regions such as intrinsically disordered regions may be more favorable for co-encoding (*19*), we explored different overlap positions and weights (fig. S4), including cases where one protein sequence was kept wild-type. Among 20 redesigned biosynthetic proteins, eight could rescue the growth of corresponding auxotrophic *Escherichia coli* bacteria in minimal media (fig. S5, A and B, and table S2). In one design, DE14, CAMEOS identified multiple optimal regions for overlapping, as shown by the ability to encode sequences similar to two essential proteins into separate regions of the functional cysteine biosynthesis gene *cysJ* (fig. S5C). The wild-type copy was knocked out in the cell to assess the co-encoded essential gene. In another design, DE2, we successfully generated a chromosomal deletion of the essential *acpP* gene in a strain containing overlapping *ilvA* and *acpP* genes. This DE2 construct encoded a redesigned protein sequence of IlvA (threonine deaminase, used in isoleucine biosynthesis) and a wild-type sequence of ACP (acyl-carrier protein, involved in essential fatty acid biosynthesis). We removed any plasmid-associated effects by genomically integrating this *ilvA-acpP* construct into a strain with deleted wild-type *ilvA* and *acpP* (fig. S6, A to C). The resulting strain (IA-1) exhibited isoleucine prototrophy at a wild-type level (Fig. 1D and fig. S6D), indicating functional activity of both biosynthetic and essential proteins.

To verify that sequence entanglement could restrict accumulation of mutations, we performed saturation mutagenesis on the genomic *ilvA-acpP* locus (Fig. 2A). Fitness effects for the first 30 overlapping codons of *ilvA* were assessed using oligo-recombineering and sequencing (*20*). Many mutations, especially in the beginning of the overlapping region, were found to incur a fitness defect (Fig. 2B and fig. S7). Although *ilvA* is not essential, 12.5% of *ilvA* mutations caused severe growth defects (decrease in growth rate by a factor of >10) (Fig. 2C). In contrast, mutagenesis of *ilvA-acpP* in a control strain (IA-2) that has an additional wild-type copy of *acpP* produced mutants with practically no growth defects (Fig. 2, B and C), which suggests that entanglement with the essential *acpP* gene renders the recoded *ilvA* gene sensitive to mutations. In the entangled sequence, a reduction in the degeneracy of codons was also observed, with 32% of synonymous codons in *ilvA* exhibiting high variability in their fitness impact because many were now deleterious (Fig. 2D). For example, the six leucine (L) codons had highly variable fitness effects across most of the overlap region analyzed (87%) (Fig. 2, B and D). Serial passaging of IA-1 and IA-2 showed that the overlap sequence remained unchanged in IA-1 after 150 generations, whereas mutations appeared in IA-2 by generation 50 (fig. S8). Together, these results demonstrate that a gene that overlaps with an essential gene is more evolutionarily constrained.

Because functional testing of in silico designs is a bottleneck, we devised a high-throughput synthesis and selection strategy to experimentally evaluate thousands of CAMEOS solutions. We used *cysJ*, a flavin sulfite reductase subunit (CysJ) for cysteine biosynthesis, and *infA*, the essential translation initiation factor–1 (IF1), as a test case by designing 7500 unique *cysJ-infA* overlapping solutions (Fig. 3A). CAMEOS designs were synthesized as a pool of 230–base pair

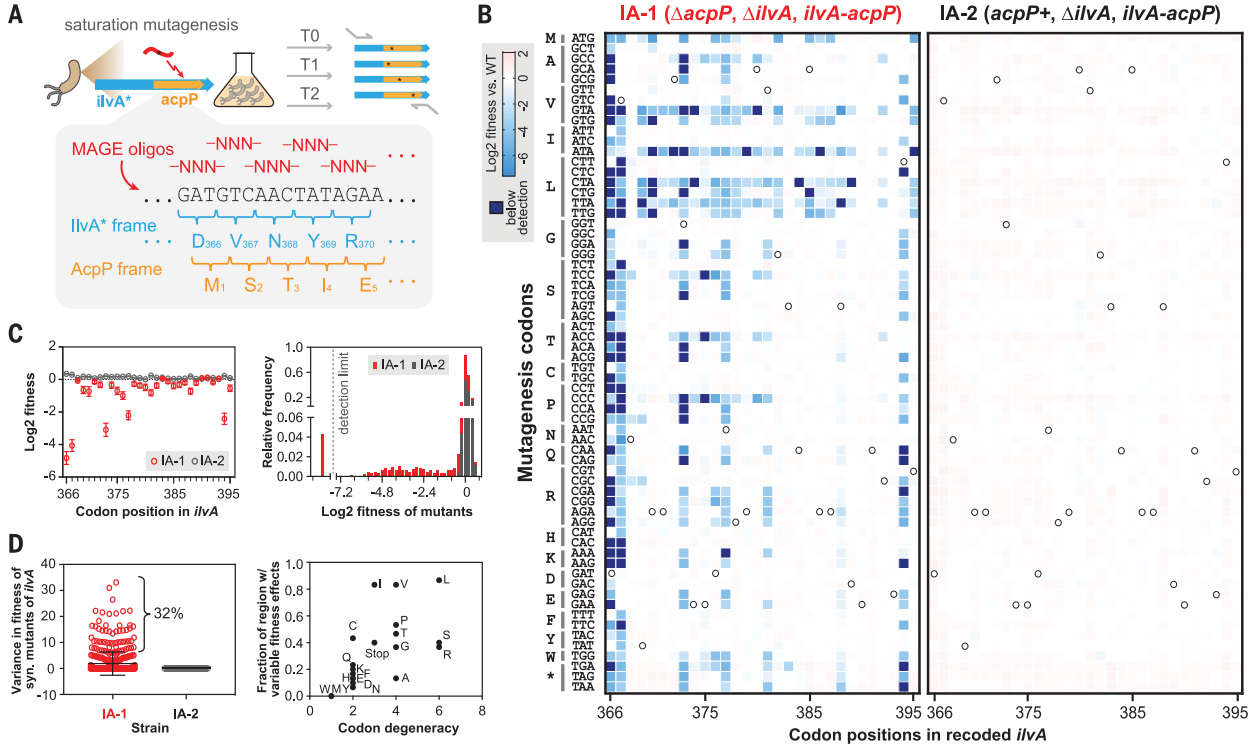


**Fig. 2. Sequence entanglement alters the protein fitness landscape.** (**A**) Saturation mutagenesis of the *ilvA-acpP* overlap region. (**B**) Fitness of all single-codon mutants of *ilvA-acpP* in strains IA-1 and IA-2. The *x* axis represents codon positions in the *ilvA*-coding frame; the *y* axis represents 64 mutagenized codons. White circles indicate wild-type *ilvA* codons. Heat map shows mutational fitness impact; dark blue indicates severe effects. (**C**) Left: Average fitness of all single codon mutants at each codon position in IA-1 (red) and IA-2 (gray) strains. Error bars denote SEM of the 64 codon mutants. Right: Distribution of fitness of all mutants in IA-1 and IA-2. (**D**) Left: Variance in fitness between synonymous mutants at each *ilvA* codon position. As shown, 32% of the codon substitutions in IA-1 have variances in fitness beyond the 95% confidence interval of IA-2 variances. Right: Fraction of the overlap region with highly variable fitness between synonymous codons for each amino acid plotted against its codon degeneracy. Amino acid abbreviations: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr; *, Stop.

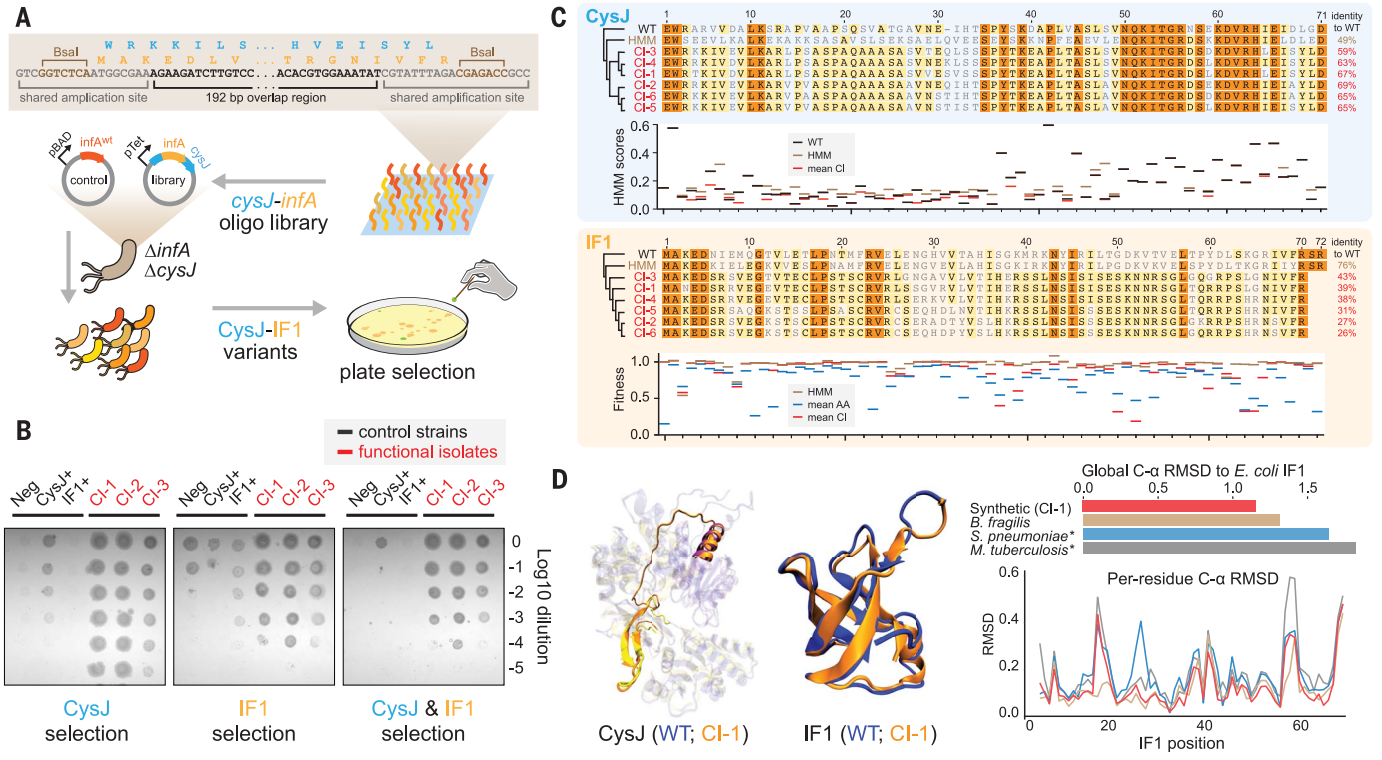


**Fig. 3. High-throughput experimental evaluation of CAMEOS designs.** (**A**) Diagram of selection platform to identify functional *cysJ-infA* variants. (**B**) Growth of *cysJ-infA* variants and controls under different plate selections. CI-1, CI-2, and CI-3 are different isolates encoding functional *cysJ-infA* variants. Control strains: Neg, cells containing pH9-cysJ$^{neg}$-infA$^{neg}$ with nonsense mutations in both *cysJ* and *infA* reading frames; CysJ$^+$, cells expressing wild-type CysJ (from pH9-cysJ$^{wt}$); IF1$^+$, cells expressing wild-type IF1 (from pH9-cysJ-infA$^{wt}$). Cultures of each strain were spotted on selection plates over serial dilutions. (**C**) Multiple sequence alignment of IF1 and CysJ encoded by six functional *cysJ-infA* variants (CI-1 to CI-6). Colored shading represents different degrees of sequence identity: orange, 100%; yellow, >60%. Shown in panels below the sequence alignments are average fitnesses of IF1 mutants at every position based on saturation mutagenesis data (mean confidence interval) and single-residue substitution scores of CysJ based on the HMM model (mean confidence interval). WT, *E. coli* wild-type sequence; HMM, consensus sequence from HMM models; mean AA, average fitness of 20 amino acids. (**D**) Left: Structural modeling of CI-1 proteins shows concordance between predicted (orange) and crystal (CysJ, yellow/purple; IF1, blue) structures. The wild-type CysJ structure was generated by concatenating separately crystalized domains (yellow, purple). Right: Global (top) and per-residue (bottom) RMSD comparing IF1 from CI-1 with an ortholog model (*B. fragilis*) and other crystal structures (denoted by an asterisk: *S. pneumoniae*, *M. tuberculosis*) shows structural similarity to the *E. coli* IF1.

oligonucleotides (fig. S9 and table S6), cloned into a pH9-cysJ-infA$^{entry}$ plasmid (fig. S10), and transformed into a Δ*cysJ-infA* strain (CI-Δ) for selection of functional CysJ and IF1 variants in MOPS dropout media (fig. S10) (*15*). Accordingly, CysJ and IF1 positive controls were only viable under their respective selective conditions (fig. S11). Plating of the variant library under double selection for both CysJ and IF1 function produced hundreds of colonies. We picked a number of colonies and reverified their phenotypes clonally (Fig. 3B) (*15*). Six unique sequences (CI-1 to CI-6) encoding different functional CysJ and IF1 variants were identified (Fig. 3C). Notably, all clones exhibited higher homology to CysJ (mean residue identity ~65%) than to IF1 (~34%) as well as lower homology to wild-type *E. coli* CysJ and IF1 than most natural variants (fig. S12). Surprisingly, functional IF1 variants contained residues that were deleterious as single-residue substitutions (*20*) (Fig. 3C). Analysis of the HMM likelihood scores of CysJ variants also revealed many single residues with predicted low fitness, which suggests that epistasis was exploited in our redesigned sequences. Structural modeling (*21*) of the CI-1 pair showed good structural alignment with wild-type CysJ and IF1, comparable to other natural orthologs (Fig. 3D). A large-scale computational analysis to overlap 119 essential with 49 biosynthetic *E. coli* proteins (*15*), which yielded ~5.8 million designs, showed that 531 of 5831 possible pairs (~9%) had pseudo-likelihood scores better than those of the experimentally verified *cysJ-infA* pairs (fig. S13). Accordingly, we estimate that 80% of these biosynthetic proteins could be encoded with at least one essential protein (fig. S14). Taken together, these results highlight that functional overlaps may exist for many gene pairs, which can be designed and evaluated at high throughput.

Finally, we hypothesized that sequence entanglement can also be used to generate biocontainment barriers that suppress unintended horizontal gene transfer (HGT). If a toxin gene is embedded in a gene of interest (GOI), recipients that lack the antitoxin would be killed when the co-acquired toxin is expressed (Fig. 1A). We thus tested designs of various bacterial toxins embedded in the *ilvA* gene and found *ilvA-ccdB* to be the best overlap pairing (Fig. 4, A and B, and fig. S15) (*15*). We then transformed conjugation-competent donors (D1, D2) expressing the antitoxin *ccdA* with plasmids carrying either *ilvA-ccdB* (T1) or *ilvA-ccdB$^{stop}$* (T2, containing a nonsense *ccdB* mutation) and incubated them with CcdA$^-$ (R1) or CcdA$^+$ (R2) recipients. As expected, CcdA$^+$ recipients acquired *ilvA-ccdB* or *ilvA-ccdB$^{stop}$* plasmids from donors at similarly high efficiencies. In contrast, CcdA$^-$ recipients acquired *ilvA-ccdB* at a much reduced frequency (by a factor of >2700) relative to the toxin-null *ilvA-ccdB$^{stop}$* control (Fig. 4C), thus demonstrating HGT suppression by CcdB-mediated killing. Interestingly, recipients that did acquire *ilvA-ccdB* remained auxotrophic for isoleucine (Fig. 4D). Accordingly, we found *ilvA-ccdB* mutations in these escapees that inactivated both *ilvA* and *ccdB* function. These results show that synthetic entanglement with a

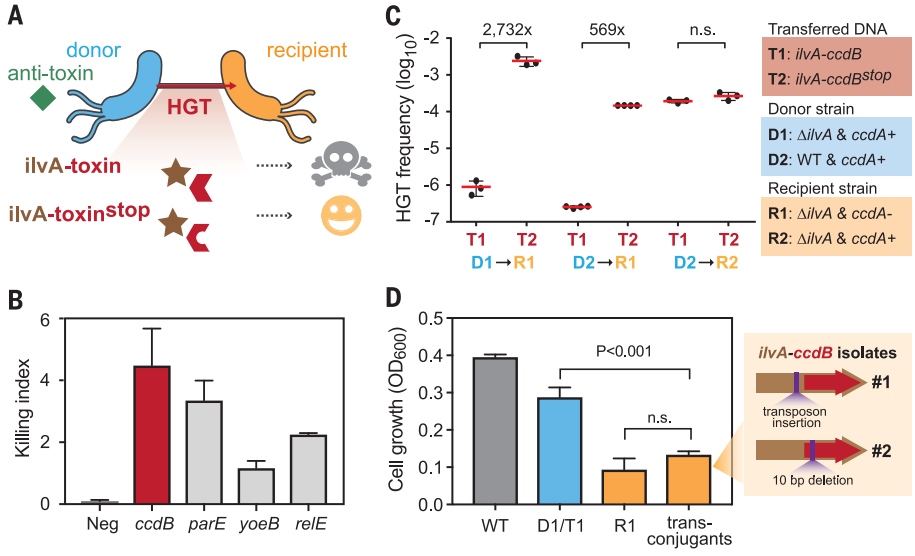


**Fig. 4. Entanglement with a toxin limits HGT.** (**A**) A toxin or an inactivated variant (toxin$^{stop}$) is overlapped with *ilvA* to assess HGT. (**B**) The killing index is shown for four *ilvA-toxin* gene pairs. Neg: parental strain without the *ilvA-toxin* construct. Data are means ± SEM from three independent experiments. (**C**) Efficiency of HGT of pHT plasmids (T1/T2) between different donor (D1/D2) and recipient (R1/R2) strains. Data are means ± SEM from three or four independent experiments. (**D**) Growth of R1 transconjugant isolates that acquired *ilvA-ccdB* (T1) after HGT, compared to the D1 donor strain (*ilvA*$^+$) and the R1 recipient strain (Δ*ilvA*) before HGT in M9 minimal media. Data are means ± SEM of growth measurements after 16 hours from three to five independent colonies and 28 R1 transconjugant isolates. Mutations identified in two R1 transconjugant isolates are illustrated at the right. Tukey's multiple comparison test was used to assess statistical significance (n.s., not significant).

toxin suppresses HGT and yields escapees that often carry nonfunctional GOI mutants.

This work describes the successful design of two full-length proteins into the same DNA under the constraints of overlap encoding. CAMEOS enables large leaps in protein space that are challenging to achieve through stepwise evolution. Our study goes beyond prior computational co-encoding efforts that used simple first-order BLOSUM substitution scores (*22*). Further improvements in MRF optimization (*23*) or explicit integration of protein structure information (*24*), along with higher-throughput gene synthesis methods (*25*), will facilitate the generation of longer overlapping genes with even higher performance. Better translation tuning may improve overlap designs, because suboptimal translation may be a failure mode. Ultimately, these advances can yield next-generation synthetic elements and circuits that will operate only in predefined settings, with greater robustness to mutations and over longer time scales.

## REFERENCES AND NOTES


1. B. G. Barrell, G. M. Air, C. A. Hutchison 3rd, *Nature* **264**, 34–41 (1976).
2. C. A. Spencer, R. D. Gietz, R. B. Hodgetts, *Nature* **322**, 279–281 (1986).
3. I. B. Rogozin *et al.*, *Trends Genet.* **18**, 228–232 (2002).
4. C. R. Sanna, W.-H. Li, L. Zhang, *BMC Genomics* **9**, 169 (2008).
5. Z. I. Johnson, S. W. Chisholm, *Genome Res.* **14**, 2268–2272 (2004).
6. M. Mizokami *et al.*, *J. Mol. Evol.* **44** (suppl. 1), S83–S90 (1997).
7. J. Choi, Z. Xu, J. H. Ou, *Mol. Cell. Biol.* **23**, 1489–1497 (2003).
8. T. Miyata, T. Yasunaga, *Nature* **272**, 532–535 (1978).
9. E. Simon-Loriere, E. C. Holmes, I. Pagán, *Mol. Biol. Evol.* **30**, 1916–1928 (2013).
10. B. Csörgo, T. Fehér, E. Tímár, F. R. Blattner, G. Pósfai, *Microb. Cell Fact.* **11**, 11 (2012).
11. F. K. Balagaddé, L. You, C. L. Hansen, F. H. Arnold, S. R. Quake, *Science* **309**, 137–140 (2005).
12. B. R. Jack *et al.*, *ACS Synth. Biol.* **4**, 939–943 (2015).
13. A. Chavez *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3669–3673 (2018).
14. J. W. Lee, C. T. Y. Chan, S. Slomovic, J. J. Collins, *Nat. Chem. Biol.* **14**, 530–537 (2018).
15. See supplementary materials.
16. S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S. I. Lee, C. J. Langmead, *Proteins* **79**, 1061–1078 (2011).
17. H. Kamisetty, S. Ovchinnikov, D. Baker, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15674–15679 (2013).
18. H. M. Salis, The ribosome binding site calculator. *Methods Enzymol.* **498**, 19–42 (2011).
19. C. Rancurel, M. Khosravi, A. K. Dunker, P. R. Romero, D. Karlin, *J. Virol.* **83**, 10719–10736 (2009).
20. E. D. Kelsic *et al.*, *Cell Syst.* **3**, 563–571.e6 (2016).
21. L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. E. Sternberg, *Nat. Protoc.* **10**, 845–858 (2015).
22. V. Opuu, M. Silvert, T. Simonson, *Sci. Rep.* **7**, 15873 (2017).
23. V. Kolmogorov, *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 1568–1583 (2006).
24. B. Kuhlman *et al.*, *Science* **302**, 1364–1368 (2003).
25. C. Plesa, A. M. Sidore, N. B. Lubock, D. Zhang, S. Kosuri, *Science* **359**, 343–347 (2018).


## ACKNOWLEDGMENTS


We thank members of the Wang lab for advice and comments and D. Baker and S. Ovchinnikov for sharing the GREMLIN source code and data (available at https://gremlin2.bakerlab.org/preds.php?db=ECOLI). **Funding:** Supported by DARPA (W911NF-15-2-0065) and the Sloan Foundation (FR-2015-65795) (H.H.W.). **Author contributions:** T.B., H.-I.H., and H.H.W. developed the initial concept; T.B. and H.-I.H. performed experiments and analyzed the results under the supervision of H.H.W. All authors wrote the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Sequencing data associated with this study are available in the National Center for Biotechnology Information Sequence Read Archive under PRJNA554669. The CAMEOS code is available at www.github.com/wanglabcumc/CAMEOS (commit a8d1ad3).


## SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/365/6453/595/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S15
Tables S1 to S9
References (*26–31*)


28 September 2018; resubmitted 21 June 2019
Accepted 15 July 2019
10.1126/science.aav5477

# Synthetic sequence entanglement augments stability and containment of genetic information in cells

Tomasz Blazejewski, Hsing-I Ho and Harris H. Wang

*Science* **365** (6453), 595-598.
DOI: 10.1126/science.aav5477

**Overlapping genes for synthetic biology**

Overlapping genes yield multiple distinct proteins when translated in alternative reading frames of the same nucleotide sequence. Blazejewski *et al.* developed a computational algorithm to predict de novo sequence entanglement and experimentally generated functional synthetic overlapping genes. When a sequence of interest was co-encoded with an essential gene in a living bacterium, its evolutionary stability substantially increased. When a gene of interest was synthetically overlapped with a toxin gene, its horizontal gene transfer frequency between bacteria was strongly suppressed. This generalizable strategy for designing, building, and testing overlapping genes helps stabilize vertical gene evolution and restrict horizontal gene flow.

*Science*, this issue p. 595