SYNTHETIC GENOMICS

Design, synthesis, and testing toward a 57-codon genome

Nili Ostrov,[1]* Matthieu Landon,[1,2,3]* Marc Guell,[1,4]* Gleb Kuznetsov,[1,5]* Jun Teramoto,[1,6] Natalie Cervantes,[1] Minerva Zhou,[7] Kerry Singh,[7] Michael G. Napolitano,[1,8] Mark Moosburner,[1] Ellen Shrock,[1] Benjamin W. Pruitt,[4] Nicholas Conway,[4] Daniel B. Goodman,[1,4] Cameron L. Gardner,[1] Gary Tyree,[1] Alexandra Gonzales,[1] Barry L. Wanner,[1,9] Julie E. Norville,[1] Marc J. Lajoie,[1]† George M. Church[1,4]†

Recoding—the repurposing of genetic codons—is a powerful strategy for enhancing genomes with functions not commonly found in nature. Here, we report computational design, synthesis, and progress toward assembly of a 3.97-megabase, 57-codon *Escherichia coli* genome in which all 62,214 instances of seven codons were replaced with synonymous alternatives across all protein-coding genes. We have validated 63% of recoded genes by individually testing 55 segments of 50 kilobases each. We observed that 91% of tested essential genes retained functionality with limited fitness effect. We demonstrate identification and correction of lethal design exceptions, only 13 of which were found in 2229 genes. This work underscores the feasibility of rewriting genomes and establishes a framework for large-scale design, assembly, troubleshooting, and phenotypic analysis of synthetic organisms.

The degeneracy of the canonical genetic code allows the same amino acid to be encoded by multiple synonymous codons (*1*). Although most organisms follow a common 64-codon template for translation of cellular proteins, deviations from this universal code found in several prokaryotic and eukaryotic genomes (*2*–*6*) have spurred the exploration of synthetic cells with expanded genetic codes.

Whole-genome synonymous codon replacement provides a mechanism to construct unique organisms exhibiting genetic isolation and enhanced biological functions. Once a codon is replaced genome-wide and its cognate transfer RNA (tRNA) is eliminated, the genomically recoded organism (GRO) can no longer translate the missing codon (*7*). Genetic isolation is achieved because DNA acquired from viruses, plasmids, and other cells would be improperly translated, which would render GROs insensitive to infection and horizontal gene transfer (fig. S1).

In addition, proteins with novel chemical properties can be explored by reassigning replaced codons to incorporate nonstandard amino acids (nsAAs), which function as chemical handles for bioorthogonal reactivity and enable biocontainment of GROs (*8*–*11*). Building on previous work that demonstrated single–stop codon replacement (*7*), we set out to explore the feasibility of multiple codon replacements genome-wide, with the aim of producing a virus-resistant, biocontained bacterium for industrial applications.

We present computational design of an *Escherichia coli* genome in which all 62,214 instances of seven different codons (5.4% of all *E. coli* codons) have been synonymously replaced, and experimental validation of 2.5 Mb (63%) of this synthetic genome (Fig. 1 and data file S1). Once completely assembled, the resulting strain ("*rE.coli-57*") would use only 57 of the canonical 64 codons (fig. S2). Although several synthetic genomes have been previously reported (*12*–*19*), a functionally

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [2]Program in Systems Biology, Harvard University, Cambridge, MA 02138, USA. [3]Ecole des Mines de Paris, Mines Paristech, Paris 75272, France. [4]Wyss Institute for Biologically Inspired Engineering, Boston, MA 02115, USA. [5]Program in Biophysics, Harvard University, Boston, MA 02115, USA. [6]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA. [7]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [8]Program in Biological and Biomedical Sciences, Harvard University, Cambridge, MA 02138, USA. [9]Department of Microbiology and Immunobiology, Harvard Medical School, Boston, MA 02115, USA.
*These authors contributed equally to this work. **†Corresponding author. Email: mlajoie@uw.edu (M.J.L.); gchurch@genetics.med.harvard.edu (G.M.C.)**

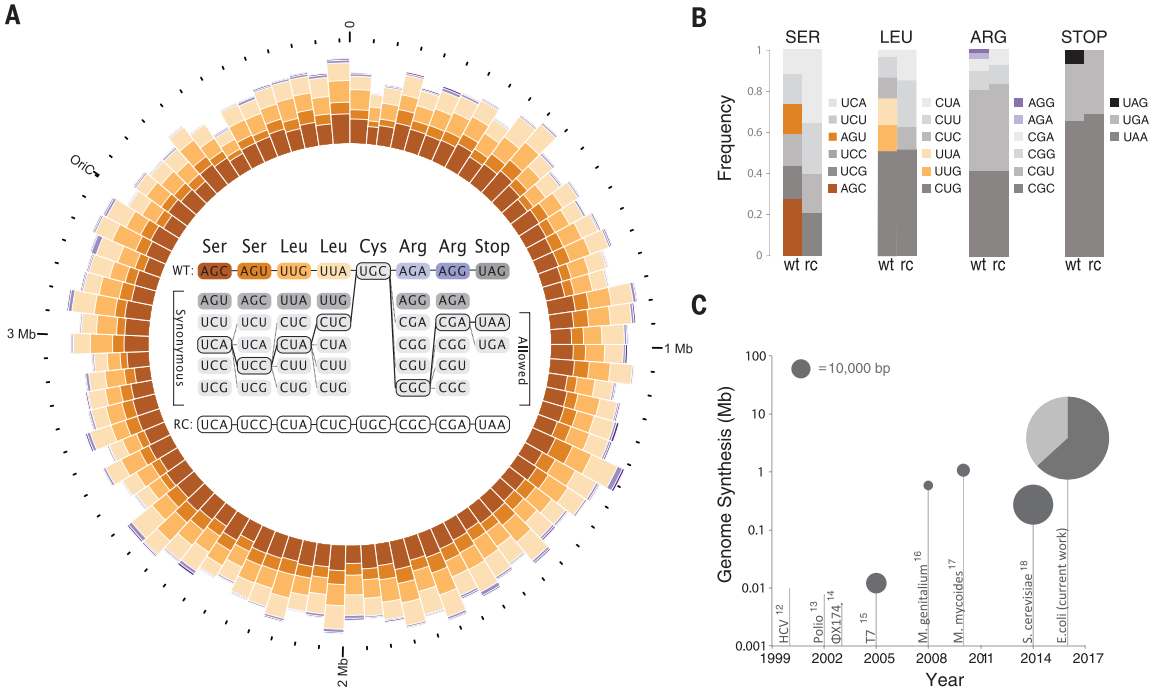

Fig. 1. A 57-codon *E. coli* genome. (**A**) The recoded genome was divided into 87 segments of ~50 kb. Codons AGA, AGG, AGC, AGU, UUA, UUG, and UAG were computationally replaced by synonymous alternatives (center). Other codons (e.g., UGC) remain unchanged. Color-coded histograms represent the abundance of the seven forbidden codons in each segment. (**B**) Codon frequencies in nonrecoded [wild-type (wt), *E. coli* MDS42] versus recoded [(rc), *rE.coli-57*] genome. Forbidden codons are colored. (**C**) The scale of DNA editing in genomes constructed by de novo synthesis. Plot area represents the number of modified base pairs compared with the parent genome. For the current work, dark gray represents percent of genome validated in vivo at time of publication (63%). HCV, hepatitis C virus; T7, bacteriophage T7; *M. genitalium*, *Mycoplasma genitalium*; and *M. mycoides*, *Mycoplasma mycoides*.

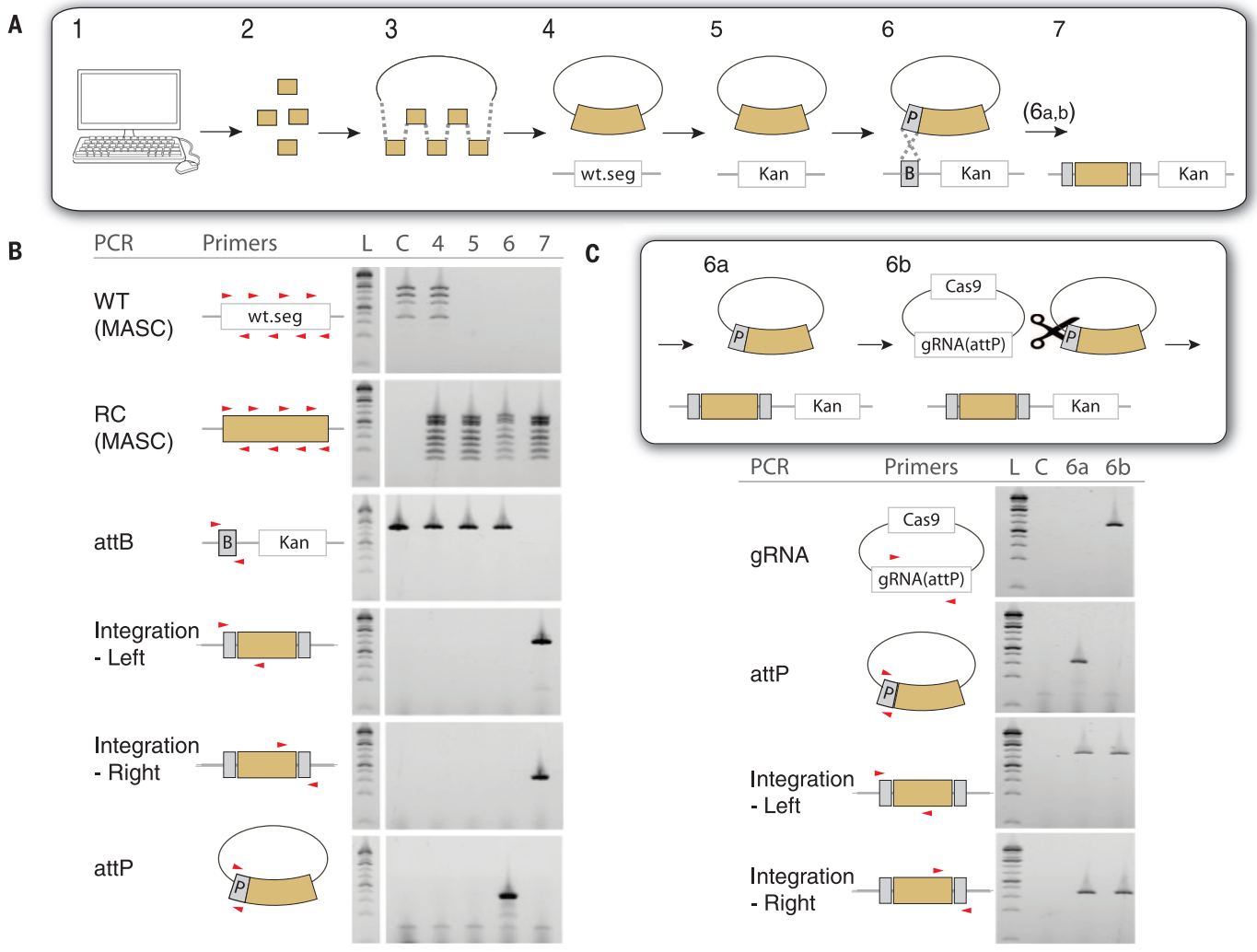

Fig. 2. Experimental strategy for recoded genome validation. (**A**) Pipeline schematics: 1) computational design; 2) de novo synthesis of 2- to 4-kb recoded fragments with 50–base pair overlap; 3) assembly of 50-kb segment (orange) in *S. cerevisiae* on a low-copy plasmid; 4) plasmid electroporation in *E. coli* (*wt.seg* is a nonrecoded chromosomal segment); 5) *wt.seg* is replaced by kanamycin cassette (Kan), such that cell viability depends solely on recoded gene expression; 6) λ-Integrase–mediated recombination of *attP* and *attB* sequences (P, episomal; B, chromosomal); 6a,b) elimination of residual vectors [see (C)]; and 7) single-copy integrated recoded segment. *attL-attR* sites shown in gray. Chromosomal deletions were performed in *E. coli* TOP10. (**B**) Polymerase chain reaction (PCR) analysis of steps 4 to 7 (L, GeneRuler 1 kb plus ladder; C, TOP10 control). Numbers correspond to schematics in (A). PCR primers shown in red (table S3). (**C**) Cas9-mediated vector elimination: Residual vector carrying recoded segment is targeted for digestion by Cas9 using *attP*-specific guide RNA (gRNA).

altered genome of this scale has not yet been explored (Fig. 1C).

Previous work suggests that codon usage alterations can affect gene expression and cellular fitness in multiple ways (*20*–*26*). However, parsing the individual impact of each codon remains difficult. Moreover, the number of modifications required to replace all instances of seven codons throughout the genome is far beyond the capabilities of current single-codon–editing strategies (*7*, *27*). Although it is possible to simultaneously edit multiple alleles using MAGE (*28*) or Cas9 (*29*), these strategies would require extensive screening with numerous oligos and likely would introduce off-target mutations (*7*, *30*). With plummeting costs of DNA synthesis, financial barriers for synthesizing entire genomes are greatly reduced, which allows for an almost unlimited number of modifications independent of biological template. Here, we developed computational and experimental tools to rapidly design and prototype synthetic organisms.

In choosing codons for replacement, UAG (stop) was selected because it was previously replaced genome-wide (*7*). AGG and AGA (Arg) are among the rarest codons in the genome, so selecting them minimizes the number of changes required. Other codons [AGC (Ser), AGU (Ser), UUG (Leu), and UUA (Leu)] were chosen such that their anticodon is not recognized as a tRNA identity element by endogenous aminoacyl-tRNA synthetases. We also considered mischarging of newly introduced tRNA by endogenous aminoacyl-tRNA synthetases upon codon reassignment. Last, we confirmed that all chosen codons are recognized by a tRNA different from that of their synonymous codons so that both codons and cognate tRNA could be eliminated (Fig. 1 and figs. S2 and S3).

In order to minimize synthesis costs and improve genome stability, we based our 57-codon genome on the reduced genome *E. coli* MDS42 (*31*). Our computational tool automated synonymous replacement of all forbidden codon occurrences in protein-coding genes while satisfying biological and technical constraints (figs. S4 and S5 and table S1). Primarily, we preserved amino acid sequences of all coding genes and adjusted DNA sequences to meet synthesis requirements (e.g., removing restriction sites, normalizing regions of extreme GC content, and reducing repetitive sequences). Alternative codons were selected to minimize disruption of biological motifs, such as ribosome binding sites (RBS) and mRNA secondary structure (*30*), and relative codon usage was conserved in order to meet translational demand (*32*, *33*). If no acceptable synonymous codon was found, the constraints were relaxed until an alternative was identified.

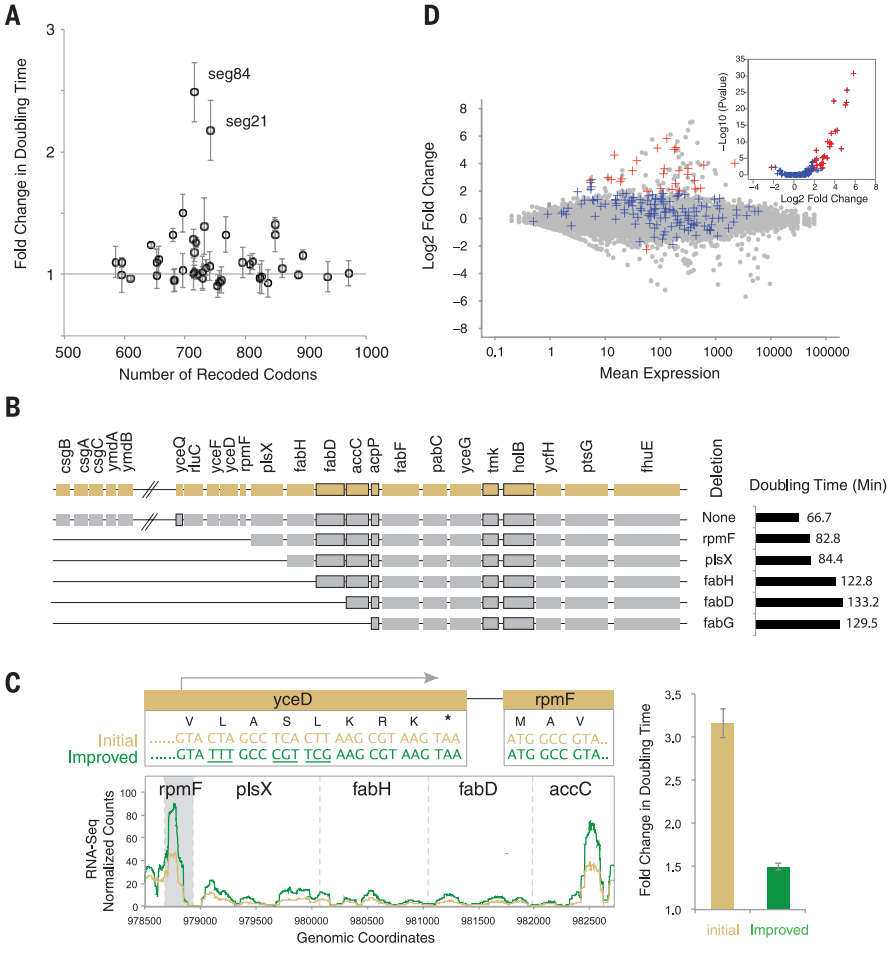

Fig. 3. Phenotypic analysis of recoded strains. (**A**) Recoded segments were episomally expressed in the absence of corresponding wild-type genes. Doubling time shown relative to nonrecoded parent strain (*35*). (**B**) Localization of fitness impairment in segment 21. Chromosomal genes (gray) were deleted to test for functional complementation by recoded genes (orange). Decrease in doubling time was observed upon deletion of *rpmF-accC* operon. Essential genes are framed. (**C**) Fine-tuning of *rpmF-accC* operon promoter resulted in increased gene expression and decrease in doubling time (normalized counts represent mean scaled sequencing depth). Orange, initial promoter; green, improved promoter. (**D**) RNA-seq analysis of 208 recoded genes (blue, segments 21, 38, 44, 46, 70). Wild-type gene expression shown in gray. Differentially expressed recoded genes shown in red (absolute log2 fold-change >2, adjusted $P < 0.01$). Fold-changes represent the difference between expression of each gene in a given strain and the average expression of the same gene in all other strains. (Inset) *P*-value distribution of recoded genes.

Forbidden codons were uniformly distributed throughout the genome and averaged ~17 codon changes per gene. Essential genes (*34*), which provide a stringent test for successful codon replacement, contained ~6.3% of all forbidden codons (3,903 of 62,214). Altogether, the recoded genome required a total of 148,955 changes to remove all instances of forbidden codons and to adjust the primary DNA sequence.

We parsed the recoded genome into 1256 synthesis-compatible overlapping fragments of 2 to 4 kb. These were used to construct 87 segments of ~50 kb each, which are convenient for yeast assembly and shuttling. Notably, intermediate 50-kb segments are also easier to troubleshoot than a full-size recoded genome or 3548 individual genes. We estimated that each segment would contain, on average, only ~1 potentially lethal recoding exception (*7*, *30*, *35*).

Carrying on average ~40 genes and ~3 essential genes, each segment was then individually tested for recoded gene functionality (fig. S4). Each segment was assembled in *Saccharomyces cerevisiae* and electroporated directly into *E. coli* on a low-copy plasmid. Subsequent deletion of the corresponding chromosomal sequence provided a stringent test for functionality of the recoded genes because errors in essential genes would be lethal (Fig. 2 and table S2).

Thus far, we performed chromosomal deletions for 2229 recoded genes (55 segments), which accounted for 63% of the genome and 53% of essential genes (fig. S6 and table S4). We thought it encouraging that 99.5% of recoded genes were found to complement wild-type genes without requiring any optimization. Moreover, the majority of chromosomally deleted strains exhibited limited fitness impairment (<10% doubling-time increase) (Fig. 3A and fig. S7).

Severe fitness impairment (>1.5-fold increase) was observed in only two strains. The causal genes were mapped by systematically removing wild-type genes, followed by measurement of strain fitness (Fig. 3, B and C, and fig. S8). We found fitness impairment in segment 21 was caused by insufficient expression of the recoded fatty acid biosynthesis operon *rpmF-accC*. Specifically, codon changes in upstream *yceD* were found to disrupt the operon promoter. Fitness was improved when *yceD* codons were altered via MAGE to preserve the overlapping promoter (Fig. 3C) (*35*). In segment 84, analysis suggested that three genes caused impairment of fitness (fig. S8), including the recoded gene *ytfP*, which contained a large deletion. Finally, RNA-sequencing (RNA-seq) analysis of 208 recoded genes suggested that the majority of genes exhibit limited change in transcription level (Fig. 3D and fig. S9) (*35*), and only 28 genes were found to be significantly differentially transcribed.

When a recoded segment failed to complement wild-type genes, it was diagnosed by making small chromosomal deletions. Notably, only 13 recoded essential genes ("design exceptions") were found that failed to support cell viability because of synonymous codon replacement (which corresponded to 9 segments) (table S4).

We chose gene *accD* as a test case to develop our troubleshooting pipeline for design exceptions (Fig. 4). First, RBS strength and mRNA folding were analyzed to pinpoint the cause of expression disruption (*22*, *23*, *26*). We further used degenerate oligos to prototype viable alternative codons (fig. S10). On the basis of these insights, a new recoded sequence was computationally generated (Fig. 4 and fig. S11) and introduced into the recoded segment via lambda Red recombineering. Viable clones were selected upon chromosomal deletion (*35*).

To further confirm adequate chromosomal expression, we integrated 11 recoded segments into the chromosome. We used *attP*-specific Cas9-mediated DNA cleavage to ablate all nonintegrated plasmids, which left a single recoded segment copy per cell. For all but one of 11 segments tested, a single copy was found to support cell viability (fig. S7). Nevertheless, we were unable to achieve a single copy of segment 32. Preliminary analysis suggested impairment of the recoded *pheMST* operon (fig. S12).

Finally, DNA sequence analysis was performed for all validated segments, showing some degree of in vivo mutations that are expected during strain engineering (table S4). Nevertheless, the average mutation rate was much lower than expected from using DNA editing methods (0.01 versus 1 mutation per codon change) (*7*). Moreover, the infrequent occurrence of reversions (25 instances in nonessential genes) speaks to the stability of the recoded genome.

It is well appreciated that without proper selection, substantial modifications to codon usage and tRNA anticodons can lead to genome instability. This could be circumvented by creating dependence on the recoded state, which also

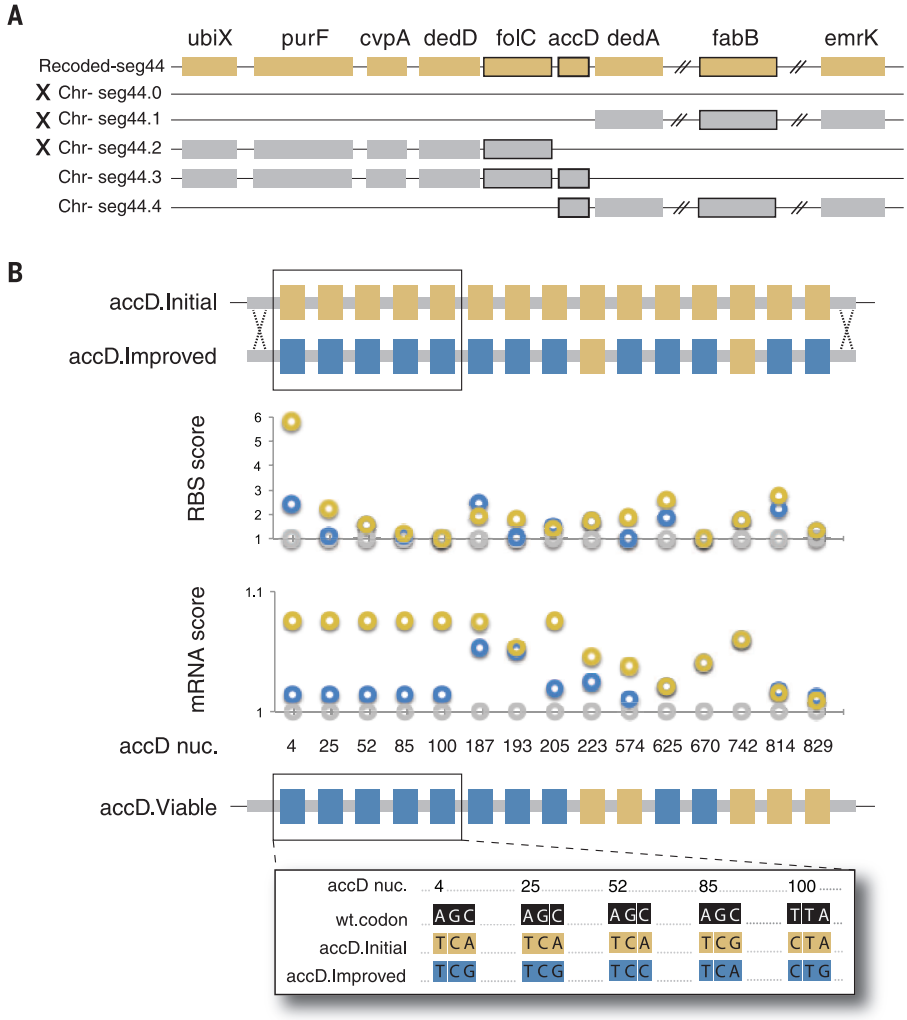

Fig. 4. Troubleshooting lethal design exceptions. (**A**) Recoded segment 44 (orange) did not support cell viability upon deletion of the corresponding chromosomal sequence (*Chr-seg44.0*). The causative recoded gene *accD* was identified by successive chromosomal deletions (*Chr-seg44.1-4*; "X," nonviable). Essential genes are framed. (**B**) λ-Red recombination was used to exchange lethal *accD* sequence (*accD.Initial*, recoded codons in orange) with an alternative recoded *accD* sequence (*accD.Improved,* alternative codons in blue). mRNA structure and RBS motif strength were calculated for both sequences. Wild type shown in gray. *accD nuc* is the first position in each recoded codon. The resulting viable sequence (*accD.Viable*) carried codons from both designs. Full sequences are provided in fig. S11. *mRNA* and RBS scores are the ratio between predicted mRNA folding energy (kcal/mol) (*37*) or predicted RBS strength (*38*) of recoded and nonrecoded codon.

provides stringent biocontainment (*9*, *36*). We previously developed a biocontained strain in which two essential genes, *adk* and *tyrS*, were altered to depend on nsAAs (*10*). Here, we confirmed that 57-codon versions of *adk* and *tyrS* were functional in vivo and that recoded and nsAA-dependent *adk* maintained fitness and provided extremely low escape rates as previously reported (fig. S13). These results suggest that the final *rE.coli-57* strain could support a similar biocontainment strategy.

As we continue toward creating a fully recoded organism, progress described herein provides crucial insights into the challenges we may encounter. The *rE.coli-57* genetic code will remain unchanged until both codons and respective tRNAs and release factors are removed (e.g., tRNAs genes *argU*, *argW*, *serV*, *leuX*, and *leuZ*, and release factor *prfA*). Only then could it be tested for novel properties, and up to four orthogonal nsAAs could be introduced.

Taken together, our results demonstrate the feasibility of radically changing the genetic code and the tractability of large-scale synthetic genome construction. A hierarchical, in vivo validation approach supported by robust design software brings the estimated total cost of constructing *rE.coli-57* to ~$1 million (table S5). Once complete, a genetically isolated *rE.coli-57* will offer a unique chassis with expanded synthetic functionality that will be broadly applicable for biotechnology.

REFERENCES AND NOTES

1. F. H. Crick, *Science* **139**, 461–464 (1963).
2. A. Ambrogelly, S. Palioura, D. Söll, *Nat. Chem. Biol.* **3**, 29–35 (2007).
3. A. Kano, Y. Andachi, T. Ohama, S. Osawa, *J. Mol. Biol.* **221**, 387–401 (1991).
4. T. Oba, Y. Andachi, A. Muto, S. Osawa, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 921–925 (1991).
5. G. Macino, G. Coruzzi, F. G. Nobrega, M. Li, A. Tzagoloff, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 3784–3785 (1979).
6. J. Ling, P. O'Donoghue, D. Söll, *Nat. Rev. Microbiol.* **13**, 707–721 (2015).
7. M. J. Lajoie *et al.*, *Science* **342**, 357–360 (2013).
8. C. C. Liu, P. G. Schultz, *Annu. Rev. Biochem.* **79**, 413–444 (2010).
9. P. Marliere, *Syst. Synth. Biol.* **3**, 77–84 (2009).
10. D. J. Mandell *et al.*, *Nature* **518**, 55–60 (2015).
11. A. J. Rovner *et al.*, *Nature* **518**, 89–93 (2015).
12. K. J. Blight, A. A. Kolykhalov, C. M. Rice, *Science* **290**, 1972–1974 (2000).
13. J. Cello, A. V. Paul, E. Wimmer, *Science* **297**, 1016–1018 (2002).
14. H. O. Smith, C. A. Hutchison 3rd, C. Pfannkoch, J. C. Venter, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 15440–15445 (2003).
15. L. Y. Chan, S. Kosuri, D. Endy, *Mol. Syst. Biol.* **1**, 0018 (2005).
16. D. G. Gibson *et al.*, *Science* **319**, 1215–1220 (2008).
17. D. G. Gibson *et al.*, *Science* **329**, 52–56 (2010).
18. N. Annaluru *et al.*, *Science* **344**, 55–58 (2014).
19. C. A. Hutchison 3rd *et al.*, *Science* **351**, aad6253 (2016).
20. G. Kudla, A. W. Murray, D. Tollervey, J. B. Plotkin, *Science* **324**, 255–258 (2009).
21. T. Tuller, Y. Y. Waldman, M. Kupiec, E. Ruppin, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 3645–3650 (2010).
22. J. B. Plotkin, G. Kudla, *Nat. Rev. Genet.* **12**, 32–42 (2011).
23. D. B. Goodman, G. M. Church, S. Kosuri, *Science* **342**, 475–479 (2013).
24. M. Zhou *et al.*, *Nature* **495**, 111–115 (2013).
25. T. E. F. Quax, N. J. Claassens, D. Söll, J. van der Oost, *Mol. Cell* **59**, 149–161 (2015).
26. G. Boël *et al.*, *Nature* **529**, 358–363 (2016).
27. F. J. Isaacs *et al.*, *Science* **333**, 348–353 (2011).
28. H. H. Wang *et al.*, *Nature* **460**, 894–898 (2009).
29. K. M. Esvelt *et al.*, *Nat. Methods* **10**, 1116–1121 (2013).
30. M. J. Lajoie *et al.*, *Science* **342**, 361–363 (2013).
31. G. Pósfai *et al.*, *Science* **312**, 1044–1046 (2006).
32. K. Temme, D. Zhao, C. A. Voigt, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 7085–7090 (2012).
33. A. H. Yona *et al.*, *Elife* **2**, e01339 (2013).
34. Y. Yamazaki, H. Niki, J. Kato, *Methods Mol. Biol.* **416**, 385–389 (2008).
35. Materials and methods are available as supplementary materials at the *Science* website.
36. S. Osawa, T. H. Jukes, *J. Mol. Evol.* **28**, 271–278 (1989).
37. N. R. Markham, M. Zuker, *Nucleic Acids Res.* **33** (Web Server), W577-W581 (2005).
38. H. M. Salis, *Methods Enzymol.* **498**, 19–42 (2011).

ACKNOWLEDGMENTS

Funding for this work was provided by U.S Department of Energy grant DE-FG02-02ER63445 and Defense Advanced Research Projects Agency grant BAA-12-64. B.L.W. was supported by NSF grant 106394; G.K. was supported by DOD NDSEG Fellowship; D.B.G. was supported by NSF Graduate Research Fellowship; E.S. was supported by the Origins of Life Initiative at Harvard University. We thank Gen9 (D. Leake, I. Saaem, and E. Nickerson), SGI-DNA (D. Gibson), and Twist (E. Leproust) for DNA synthesis and J. Aach for insightful comments. G.K., M.J.L., M.L., M.G.N., D.B.G., and G.M.C. are inventors on patent application #62350468 submitted by the President and Fellows of Harvard College. The authors declare competing financial interests: Details are available in the supplementary documents of the paper. All FASTQ sequence files have been submitted to the Sequence Read Archive (SRA) at National Center for Biotechnology Information submission number SUB1484507. Plasmids will be available under a materials transfer agreement with Addgene.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/353/6301/819/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S13
Tables S1 to S5
References (*39*–*58*)
Data File S1

29 January 2016; accepted 21 July 2016
10.1126/science.aaf3639

Editor's Summary

Recoding and repurposing genetic codons

By recoding bacterial genomes, it is possible to create organisms that can potentially synthesize products not commonly found in nature. By systematic replacement of seven codons with synonymous alternatives for all protein-coding genes, Ostrov *et al*. recoded the *Escherichia coli* genome. The number of codons in the *E. coli* genetic code was reduced from 64 to 57 by removing instances of the UAG stop codon and excising two arginine codons, two leucine codons, and two serine codons. Over 90% functionality was successfully retained. In 10 cases, reconstructed bacteria were not viable, but these few failures offered interesting insights into genome-design challenges and what is needed for a viable genome.

Science, this issue p. 819