Emergent rules for codon choice elucidated by editing rare arginine codons in *Escherichia coli*

Michael G. Napolitano[a,b,c,1], Matthieu Landon[a,b,d,e,1], Christopher J. Gregg[a,b,1], Marc J. Lajoie[a,b,f,1,2], Lakshmi Govindarajan[g], Joshua A. Mosberg[a,b,f], Gleb Kuznetsov[a,b,h], Daniel B. Goodman[a,b,i], Oscar Vargas-Rodriguez[j], Farren J. Isaacs[k], Dieter Söll[j,l], and George M. Church[a,b,2]

[a]Department of Genetics, Harvard Medical School, Boston, MA 02115; [b]Wyss Institute for Biologically Inspired Engineering, Harvard Medical School, Boston, MA 02115; [c]Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02115; [d]Systems Biology Graduate Program, Harvard Medical School, Boston, MA 02115; [e]Ecole des Mines de Paris, Mines Paristech, 75272 Paris, France; [f]Program in Chemical Biology, Harvard University, Cambridge, MA 02138; [g]National University of Singapore, School of Computing, Singapore 117417; [h]Program in Biophysics, Harvard University, Boston, MA 02115; [i]Harvard-MIT Health Sciences and Technology, Cambridge, MA 02139; [j]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06516; [k]Molecular, Cellular, Developmental and Systems Biology Institute, Yale University, New Haven, CT 06516; and [l]Department of Chemistry Yale University, New Haven, CT 06516

Edited by Tina M. Henkin, The Ohio State University, Columbus, OH, and approved July 29, 2016 (received for review April 22, 2016)

The degeneracy of the genetic code allows nucleic acids to encode amino acid identity as well as noncoding information for gene regulation and genome maintenance. The rare arginine codons AGA and AGG (AGR) present a case study in codon choice, with AGRs encoding important transcriptional and translational properties distinct from the other synonymous alternatives (CGN). We created a strain of *Escherichia coli* with all 123 instances of AGR codons removed from all essential genes. We readily replaced 110 AGR codons with the synonymous CGU codons, but the remaining 13 "recalcitrant" AGRs required diversification to identify viable alternatives. Successful replacement codons tended to conserve local ribosomal binding site-like motifs and local mRNA secondary structure, sometimes at the expense of amino acid identity. Based on these observations, we empirically defined metrics for a multidimensional "safe replacement zone" (SRZ) within which alternative codons are more likely to be viable. To evaluate synonymous and nonsynonymous alternatives to essential AGRs further, we implemented a CRISPR/Cas9-based method to deplete a diversified population of a wild-type allele, allowing us to evaluate exhaustively the fitness impact of all 64 codon alternatives. Using this method, we confirmed the relevance of the SRZ by tracking codon fitness over time in 14 different genes, finding that codons that fall outside the SRZ are rapidly depleted from a growing population. Our unbiased and systematic strategy for identifying unpredicted design flaws in synthetic genomes and for elucidating rules governing codon choice will be crucial for designing genomes exhibiting radically altered genetic codes.

codon choice | genome editing | recoded genomes

The genetic code possesses inherent redundancy (1), with up to six different codons specifying a single amino acid. Although it is tempting to approximate synonymous codons as equivalent (2), most prokaryotes and many eukaryotes (3, 4) display a strong preference for certain codons over synonymous alternatives (5, 6). Although different species have evolved to prefer different codons, codon bias is largely consistent within each species (5). However, within a given genome, codon bias differs among individual genes according to codon position, suggesting that codon choice has functional consequences. For example, rare codons are enriched at the beginning of essential genes (7, 8), and codon use strongly affects protein levels (9–11), especially at the N terminus (12). These observations suggest that codon use plays a poorly understood role in regulating protein expression. Several hypotheses attempt to explain how codon use mediates this effect, including but not limited to facilitating ribosomal pausing early in translation to optimize protein folding (13); adjusting mRNA secondary structure to optimize translation initiation or to modulate mRNA degradation; preventing ribosome stalling by coevolving with tRNA levels (6); providing a "translational ramp" for proper ribosome spacing and effective translation (14); and providing a layer of translational regulation for independent control of each gene in an operon (15). Additionally, codon use may impact translational fidelity (16), and the proteome may be tuned by fine control of the decoding tRNA pools (17). Although Quax et al. (18) provide an excellent review of how biology chooses codons, systematic and exhaustive studies of codon choice in whole genomes are lacking. Studies have only begun to probe the effects of codon choice empirically in a relatively small number of reporter genes (12, 19–22). Several important questions must be answered as a first step toward designing custom genomes exhibiting new functions: How flexible is genome-wide codon choice? How does codon choice interact with the maintenance of cellular homeostasis? What heuristics can be used to predict which codons will conserve genome function?

Replacing all essential instances of a codon in a single strain would provide valuable insight into the constraints that determine codon choice and aid in the design of recoded genomes. Although the UAG stop codon has been completely removed

Significance

This work presents the genome-wide replacement of all rare AGR (AGA and AGG) arginine codons in the essential genes of *Escherichia coli* with synonymous CGN alternatives. Synonymous codon substitutions can lethally impact noncoding function by disrupting mRNA secondary structure and ribosomal binding site-like motifs. Here we quantitatively define the range of tolerable deviation in these metrics and use this relationship to provide critical insight into codon choice in recoded genomes. This work demonstrates that genome-wide removal of AGR is likely to be possible and provides a framework for designing genomes with radically altered genetic codes.

Author contributions: M.G.N., M.L., C.J.G., M.J.L., J.A.M., F.J.I., D.S., and G.M.C. designed research; M.G.N., M.L., C.J.G., M.J.L., L.G., J.A.M., G.K., D.B.G., and O.V.-R. performed research; M.G.N., M.L., M.J.L., J.A.M., G.K., D.B.G., and O.V.-R. contributed new reagents/analytic tools; M.G.N., M.L., C.J.G., M.J.L., L.G., G.K., D.B.G., and O.V.-R. analyzed data; and M.G.N., M.L., and M.J.L. wrote the paper.

Conflict of interest statement: 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. G.M.C. is a founder of Enevolv Inc. and Gen9bio. Other potentially relevant financial interests are listed at arep.med.harvard.edu/gmc/tech.html.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The sequence reported in this paper have been deposited in the BioProject database, ncbi.nlm.nih.gov/bioproject (accession no. PRJNA298327).

[1]M.G.N., M.L., C.J.G., and M.J.L. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: gchurch@genetics.med.harvard.edu or mlajoie@uw.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1605856113/-/DCSupplemental.

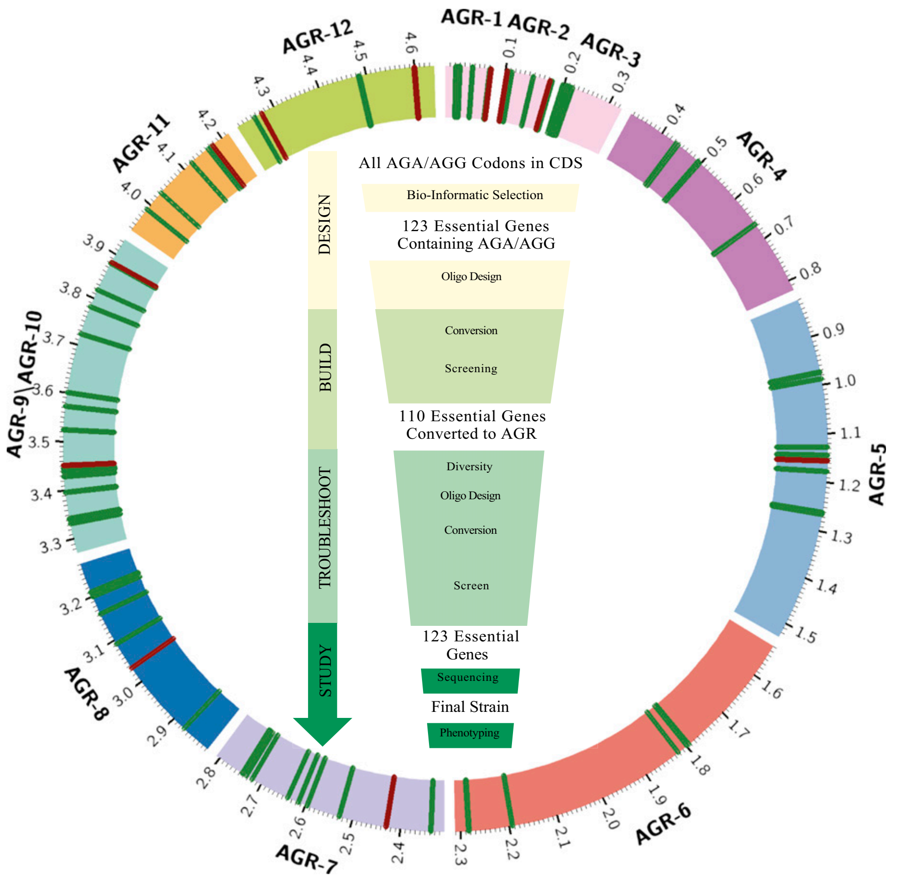

Fig. 1. Construction of strain C123. (*Inner*) Workflow used to create and analyze strain C123. The DESIGN phase involved identification of 123 AGR codons in the essential genes of *E. coli*. MAGE oligos were designed to replace all instances of these AGR codons with the synonymous CGU codon. The BUILD phase used CoS-MAGE to convert 110 AGR codons to CGU and to identify 13 AGR codons that required additional troubleshooting. The in vivo TROUBLESHOOTING phase resolved the 13 codons that could not be readily converted to CGU and identified mechanisms potentially explaining why AGR→CGU was not successful. In the STUDY phase, next-generation sequencing, evolution, and phenotyping were performed on strain C123. (*Outer*) Schematic of the C123 genome (nucleotide 0 is oriented up; numbering is according to strain MG1655). Exterior labels indicate the set groupings of AGR codons. Successful AGR→CGU conversions (110 instances) are indicated by radial green lines, and recalcitrant AGR codons (13 instances) are indicated by radial red lines.

from *Escherichia coli* (23), no genome-wide replacement of a sense codon has been reported. Although the translation function of the AGG codon has been shown to permit efficient suppression with nonstandard amino acids (24–26), AGG necessarily remains translated as Arg in each of these studies. No study has yet demonstrated that all instances of any sense codon can be removed from essential genes. These insights are crucial for unambiguously reassigning sense codon translation function.

We chose to study the rare Arg codons AGA and AGG [termed "AGR" according to International Union of Pure and Applied Chemistry (IUPAC) conventions] because the literature suggests that they are among the most difficult codons to replace and that their similarity to ribosome-binding sequences (RBSs) underlies important noncoding functions (8, 27–30). Furthermore, their sparse use (123 instances in the essential genes of *E. coli* MG1655 and 4,228 instances in the entire genome) (Table 1 and Dataset S1) made replacing all AGR instances in essential genes a tractable goal, with essential genes serving as a stringent test set for identifying any fitness impact from codon replacement (31). Additionally, recent work has shown the difficulty of directly mutating some AGR codons to other synonymous codons (25), although the authors do not explain the mechanism of failure or report successful implementation of alternative designs. We attempted to remove all 123 instances of AGR codons from essential genes by replacing them with the synonymous CGU codon. CGU was chosen to disrupt the primary nucleic acid sequence maximally (AGR→CGU). We hypothesized that this strategy would maximize design flaws, thereby revealing rules for designing genomes with reassigned genetic codes. Importantly, individual codon targets were not inspected a priori to ensure an unbiased empirical search for design flaws.

Table 1. Summary of AGR codons changed by location in the genome, and failure rates by pool

AGR pool	No. AGR codons	No. successful	No. failed	% successful
AGR.1	11	10	1	91
AGR.2	12	10	2	83
AGR.3	10	10	0	100
AGR.4	7	7	0	100
AGR.5	14	13	1	93
AGR.6	8	8	0	100
AGR.7	13	11	2	85
AGR.8	9	8	1	89
AGR.9	10	9	1	90
AGR.10	13	12	1	92
AGR.11	7	6	1	86
AGR.12	9	6	3	67
Total	123	110	13	89

Colors coordinate with Fig. 1 and Fig. S10 colors for ease of use.

Results

To construct this modified genome, we used coselection multiplex automatable genome engineering (CoS-MAGE) (32, 33) to create an *E. coli* strain (C123) with all 123 AGR codons removed from its essential genes (see Fig. 1*A* and Dataset S1 for a complete list of AGR codons in essential genes). CoS-MAGE leverages Lambda Red-mediated recombination (34, 35) and exploits the linkage between a mutation in a selectable allele (e.g., *tolC*) to nearby edits of interest (e.g., AGR conversions), thereby enriching for cells with those edits (Fig. S1). To streamline C123 construction, we chose to start with *E. coli* strain EcM2.1, which was previously optimized for efficient Lambda Red-mediated genome engineering (33, 36). Using CoS-MAGE on EcM2.1 improves allele replacement frequency by 10-fold over MAGE in nonoptimized strains but performs optimally when all edits are on the same replichore and within 500 kb of the selectable allele (33). To accommodate this requirement, we divided the genome into 12 segments containing all 123 AGR codons in essential genes. A *tolC* cassette was moved around the genome to enable CoS-MAGE in each segment, allowing us to prototype each set of AGR→CGU mutations rapidly across large cell populations in vivo. (Please see *General Replacement Strategy* and *Troubleshooting Strategy* in *Materials and Methods* for a more detailed discussion). Of the 123 AGR codons in essential genes, 110 could be changed to CGU by this process (Fig. 1), revealing considerable flexibility of codon use for most essential genes. The frequency of allele replacement (in this case, AGR→CGU codon substitution) varied widely across these 110 permissive codons, with no clear correlation between the frequency of allele replacement and the normalized position of the AGR codon in a gene (Fig. 2*A*).

The remaining 13 AGR→CGU mutations were not observed, suggesting codon substitution frequency below our detection limit of 1% of the bacterial population (*Materials and Methods* and Dataset S2). These "recalcitrant codons" were assumed to be deleterious or nonrecombinogenic and were triaged into a troubleshooting pipeline for further analysis (Fig. 1). Interestingly, all except 1 of the 13 recalcitrant codons were colocalized near the termini of their respective genes, suggesting the importance of codon choice at these positions: seven were at most 30 nt downstream of the start codon, and five were at most 30 nt upstream of the stop codon (Fig. 2*A*, *Lower* and Dataset S3). Because of our unbiased design strategy, we anticipated that several AGR→CGU mutations would present obvious design flaws, such as introducing nonsynonymous mutations (two instances) or RBS disruptions (four instances) in overlapping genes. For example, *ftsI*_AGA1759 overlaps the second and third codons of *murE*, an essential gene, introducing a missense mutation (*murE* D3V) that may impair fitness. Replacing *ftsI*_AGA with CGA successfully replaced the forbidden AGA codon while conserving the primary amino acid sequence of MurE with a minimal impact on fitness (Fig. 3*A* and Dataset S2). Similarly, *holB*_AGA4 overlaps the upstream essential gene *tmk*, and replacing AGA with CGU converts the *tmk* stop codon to Cys, adding 14 amino acids to the C terminus of *tmk*. Although some C-terminal extensions are well tolerated in *E. coli* (37), extending *tmk* appears to be deleterious. We successfully replaced *holB*_AGA with CGC by inserting three nucleotides comprising a stop codon before the *holB* start codon. This insertion reduced the *tmk*/*holB* overlap and preserved the coding sequences of both genes (Fig. S2*A*).

Additionally, the four remaining C-terminal failures included AGR→CGU mutations that disrupt RBS motifs belonging to downstream genes (*secE*_AGG376 for *nusG*, *dnaT*_AGA532 for *dnaC*, and *folC*_AGAAGG1249,1252 for *dedD*, the last constituting two codons). Both *nusG* and *dnaC* are essential, suggesting that replacing AGR with CGU in *secE* and *dnaT* lethally disrupts translation initiation and thus the expression of the overlapping *nusG* and *dnaC* (Fig. 3*B* and Fig. S2*B*). Although *dedD* is annotated as nonessential (31), we hypothesized that replacing the AGR with CGU in *folC* disrupted a portion of *dedD* that is essential to the survival of EcM2.1 (*E. coli* K-12). In support of this hypothesis, we were unable to delete the 29 nucleotides of *dedD* that were not deleted by Baba et al. (31) and did not overlap with *folC*, suggesting that this sequence is essential in our strain. The unexpected failure of this conversion highlights the challenge of predicting design flaws even in well-annotated organisms. Consistent with our observation that disrupting these RBS motifs underlies the failed AGR→CGU conversions, we overcame all four design flaws by selecting codons that conserved RBS strength, including a nonsynonymous (Arg→Gly) conversion for *secE*.

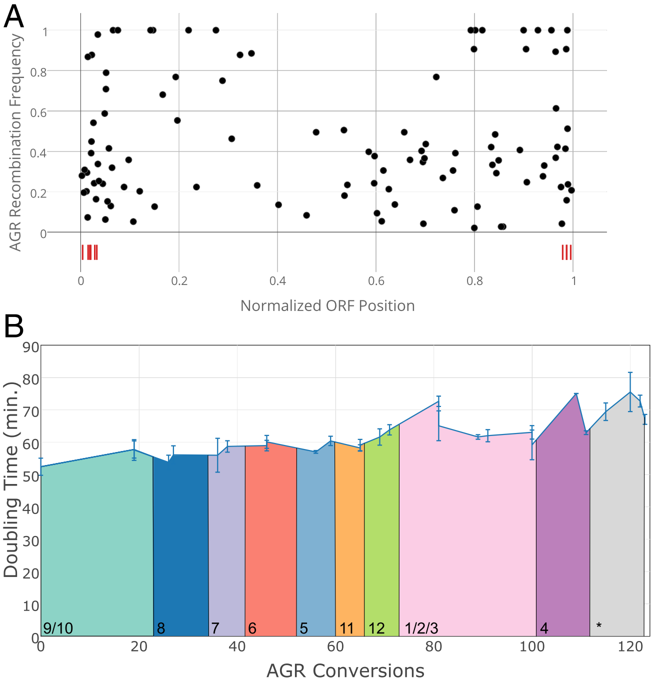

Fig. 2. Analysis of attempted AGR→CGU replacements. (*A*) AGR recombination frequency (mascPCR, n = 96 clones per cell population) was plotted versus the normalized ORF position (residue number of the AGR codon divided by the total length of the ORF). Failed AGR→CGU conversions are indicated by vertical red lines below the x axis. (*B*) Doubling time of strains in the C123 lineage in LB^L medium at 34 °C was determined in triplicate on a 96-well plate reader. Colored bars indicate the set of codons under construction when a doubling time was determined (coloring based on Fig. 1). Each data point represents a different stage of strain construction. Alternative codons were identified for 13 recalcitrant AGR codons in our troubleshooting pipeline, the optimized replacement sequences were incorporated into the final strain (gray section at right, labeled with an asterisk), and the resulting doubling times were measured. Error bars represent SEM in doubling time from at least three replicates of each strain.

These lessons, together with previous observations that ribosomes pause during translation when they encounter RBS motifs in coding DNA sequences (20), provided key insights into the N-terminal AGR→CGU failures. Three of the N-terminal failures (*ssb*_AGA10, *dnaT*_AGA10, and *prfB*_AGG64) had RBS-like motifs that were either disrupted or created by CGU replacement. Although *prfB*_AGG64 is part of the RBS motif that triggers an essential frameshift mutation in *prfB* (21, 38, 39), pausing motif-mediated regulation of *ssb* and *dnaT* expression has not been reported. Nevertheless, ribosomal pausing data (20) showed that ribosomal occupancy peaks are present directly downstream of the AGR codons for *ssb* and are absent for *dnaT* (Fig. S3); meanwhile, unsuccessful CGU mutations were predicted to weaken the RBS-like motif for *prfB* and *ssb* and to strengthen the RBS-like motif for *dnaT* (Fig. 3*C* and Fig. S2*C*), suggesting a functional relationship

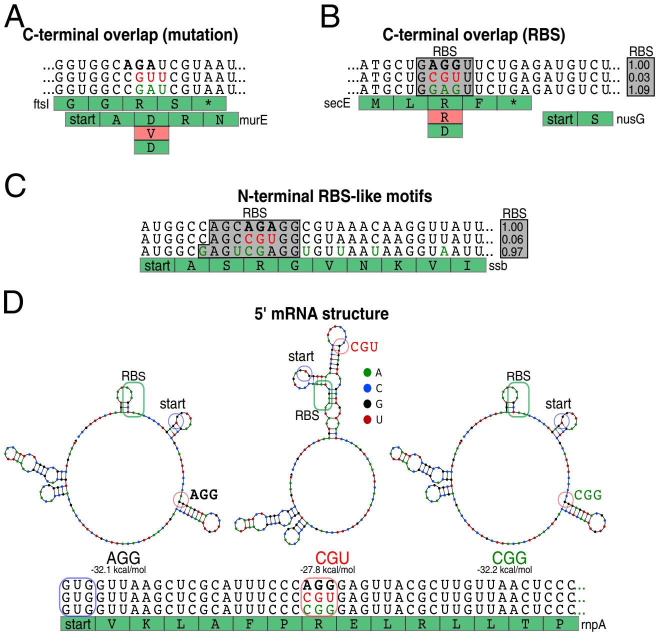
A
C-terminal overlap (mutation)
...GGUGGCAGAUCGUAAU...
...GGUGGCCGUUCGUAAU...
...GGUGGCCGAUCGUAAU...
ftsI
G G R S *
start A D R N
murE
V
D
B
C-terminal overlap (RBS)
RBS
...ATGCUGAGGUUCUGAGAUGUCU...
...ATGCUGCGUUUCUGAGAUGUCU...
...ATGCUGGAGUUCUGAGAUGUCU...
RBS
1.00
0.03
1.09
secE
M L R F *
R
D
start S
nusG
C
N-terminal RBS-like motifs
RBS
AUGGCCAGCAGAGGCGUAAACAAGGUUAUU...
AUGGCCAGCCGUGGCGUAAACAAGGUUAUU...
AUGGCGAGUCGAGGUGUUAAUAAGGUAAUU...
RBS
1.00
0.06
0.97
start A S R G V N K V I
ssb
D
5' mRNA structure
RBS
start
AGG
start
CGU
RBS
A
C
G
U
RBS
start
CGG
AGG
-32.1 kcal/mol
CGU
-27.8 kcal/mol
CGG
-32.2 kcal/mol
GUGGUUAAGCUCGCAUUUCCCAGGGAGUUACGCUUGUUAACUCCC...
GUGGUUAAGCUCGCAUUUCCCCGUGAGUUACGCUUGUUAACUCCC...
GUGGUUAAGCUCGCAUUUCCCCGGGAGUUACGCUUGUUAACUCCC...
start V K L A F P R E L R L L T P
rnpA

Fig. 3. Examples of failure mechanisms for four recalcitrant AGR replacements. Wild-type AGR codons are indicated by bold black letters, design flaws are indicated by red letters, and optimized replacement genotypes are indicated by green letters. (*A*) The genes *ftsI* and *murE* overlap with each other. An AGA→CGU mutation in *ftsI* would introduce a nonconservative Asp3Val mutation in *murE*. The amino acid sequence of *murE* was preserved by using an AGA→CGA mutation. (*B*) Gene *secE* overlaps with the RBS for the downstream essential gene *nusG*. An AGG→CGU mutation is predicted to diminish the RBS strength by 97% (53). RBS strength is preserved by using a nonsynonymous AGG→GAG mutation. (*C*) Gene *ssb* has an internal RBS-like motif shortly after its start codon. An AGG→CGU mutation would diminish the RBS strength by 94%. RBS strength is preserved by using an AGA→CGA mutation combined with additional wobble mutations indicated by green letters. (*D*) Gene *rnpA* has a defined mRNA structure that would be changed by an AGG→CGU mutation. The original RNA structure is preserved by using an AGG→CGG mutation. The RBS (green), start codon (blue), and AGR codon (red) are annotated with like-colored boxes on the predicted RNA secondary structures.

between RBS occupancy and cell fitness. Consistent with this hypothesis, successful codon replacements from the troubleshooting pipeline conserve predicted RBS strength compared with the large predicted deviation caused by unsuccessful AGR→CGU mutations (Fig. 4, y axis and comparison of orange asterisks and green dots). Interestingly, attempts to replace *dnaT*_AGA10 with either CGN or NNN failed; only by manipulating the wobble position of surrounding codons and conserving the Arg amino acid could *dnaT*_AGA10 be replaced (Fig. S2*C*). These wobble variants appear to compensate for the increased RBS strength caused by the AGA→CGU mutation: RBS motif strength with wobble variants deviated eightfold from the unmodified sequence, whereas RBS motif strength for AGA→CGU alone deviated 27-fold.

To understand better the several remaining cases of N-terminal failure that did not exhibit considerable deviations in RBS strength (*rnpA*_AGG22, *ftsA*_AGA19, *frr*_AGA16, and *rpsJ*_AGA298), we examined other potential nucleic acid determinants of protein expression. Based on the observation that the mRNA secondary structure near the 5′ end of ORFs strongly impacts protein expression (12), we found that these four remaining AGR→CGU mutations changed the predicted folding energy and structure of the mRNA near the start codon of target genes (Fig. 3*D* and Fig. S4). Successful codon replacements obtained from degenerate MAGE oligos reduced the disruption of the mRNA secondary structure compared with CGU (Fig. 4, green dots). For example, *rnpA* has a predicted mRNA loop near its RBS and start codon that relies on base pairing between both guanines of the AGG codon to nearby cytosines (Fig. 3*D* and Fig. S5*A*). Importantly, only AGG22CGG was observed out of all attempted *rnpA* AGG22CGN mutations, and the fact that only CGG preserves this mRNA structure suggests that it is physiologically important (Fig. 3*D* and Fig. S5 *B* and *C*). In support of this notion, we successfully introduced an *rnpA* AGG22CUG mutation (Arg→Leu) only when we changed the complementary nucleotides in the stem from CC (base pairs with AGG) to CA (base pairs with CUG), thus preserving the natural RNA structure (Fig. S5*D*) while changing both RBS motif strength and amino acid identity. Our analysis of all four optimized gene sequences showed reduced deviation in computational mRNA folding energy [computed with UNAFold (40)] compared with the unsuccessful CGU mutations (Fig. 4, *x* axis, orange asterisks, and green dots). Similarly, the predicted mRNA structure [computed with different mRNA folding software, NUPACK (41)] for these genes was strongly changed by CGU mutations and was corrected in our empirically optimized solutions (Fig. S4).

Troubleshooting these 13 recalcitrant codons revealed that mutations causing large deviations from natural mRNA folding energy or RBS strength are associated with failed codon substitutions. By calculating these two metrics for all attempted AGR→CGU mutations, we empirically defined a safe replacement zone (SRZ) within which most CGU mutations were tolerated (Fig. 4, shaded area). The SRZ is defined as the largest multidimensional space that contains none of the AGR→CGU failures associated with mRNA folding energy or RBS strength (Fig. 4, red asterisks). It comprises deviations in mRNA folding energy of less than 10% with respect to the natural codon and deviations in RBS-like motif scores of less than a half log with respect to the natural codon, providing a quantitative guideline for codon substitution. Notably, the optimized solution used to replace the 13 recalcitrant codons always exhibited reduced deviation for at least one of these two parameters compared with the deviation seen with a CGU mutation. Furthermore, solutions to the 13 recalcitrant codons overlapped almost entirely with the empirically defined SRZ. These results suggest that computational predictions of mRNA folding energy and RBS strength can be used as a first approximation to predict whether a designed mutation is likely to be viable. Developing *in silico* heuristics to predict problematic alleles streamlines the use of in vivo genome engineering methods such as MAGE to identify viable replacement codons empirically. Therefore, these heuristics reduce the search space required to redesign viable genomes, raising the prospect of creating radically altered genomes exhibiting expanded biological functions.

Once we had identified viable replacement sequences for all 13 recalcitrant codons, we combined the successful 110 CGU conversions with the 13 optimized codon substitutions to produce strain C123, in which all 123 AGR codons have been removed from all of its annotated essential genes. C123 then was sequenced to confirm AGR removal and analyzed using Millstone, a publicly available genome resequencing analysis pipeline (https://github.com/churchlab/millstone). Two spontaneous AAG (Lys) to AGG (Arg) mutations were observed in the essential genes *pssA* and *cca*. Although attempts to revert these mutations to AAG were unsuccessful—perhaps suggesting functional compensation—we were able to replace them with CCG (Pro) in *pssA* and CAG (Gln) in *cca* using degenerate MAGE oligos. The resulting strain, C123a, is the first strain completely devoid of AGR codons in its annotated essential genes (https://github.com/churchlab/agr_recoding) (Dataset S4). Although some AGR codons in nonessential genes could prove unexpectedly difficult to change, our success in replacing all 123 instances of AGR codons in essential genes provides strong evidence that the remaining 4,105 AGR codons can be completely removed from the *E. coli* genome, permitting the unambiguous reassignment of AGR translation function (23).

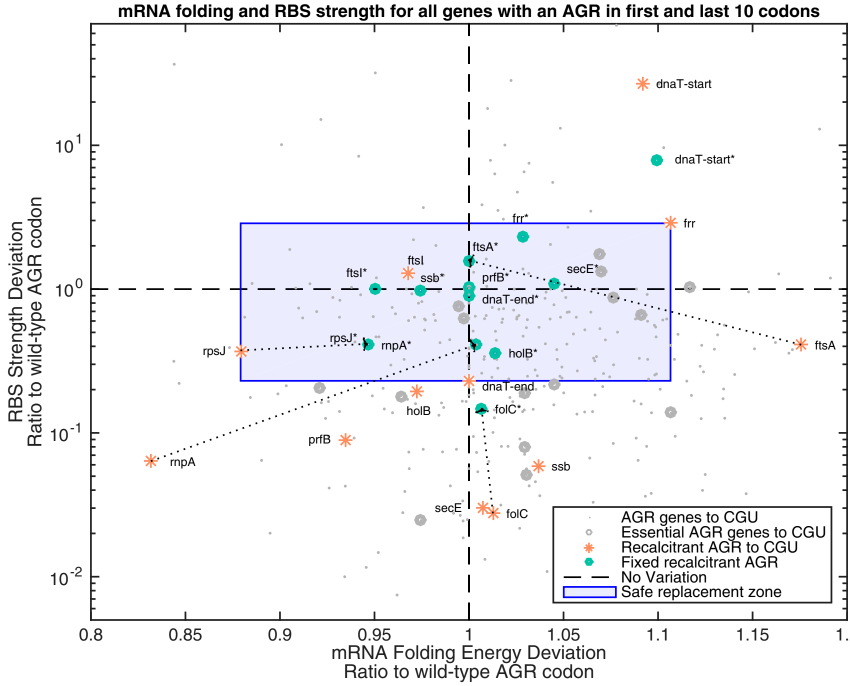

Fig. 4. RBS strength and mRNA structure predict synonymous mutation success. Scatter plot showing predicted RBS strength [*y* axis, calculated with the Salis RBS calculator (53)] versus deviations in mRNA folding [*x* axis, calculated at 37 °C by the UNAFold calculator (40)]. Small gray dots represent nonessential genes in *E. coli* MG1655 that have an AGR codon within the first 10 or last 10 codons. Large gray dots represent successful AGR→CGU conversions in the first 10 or last 10 codons of essential genes. Orange asterisks represent unsuccessful AGR→CGU mutations (recalcitrant codons) in essential genes. Green dots represent optimized solutions for these recalcitrant codons. The SRZ (blue-shaded region) is an empirically defined range of mRNA folding and RBS strength deviations, based on the successful AGR→CGU replacement mutations observed in this study. Most unsuccessful AGR→CGU mutations (orange asterisks) cause large deviations in RBS strength or mRNA structure that are outside the SRZ. The genes *holB* and *ftsI* are two notable exceptions because their initial CGU mutations caused amino acid changes in overlapping essential genes. Gene *folC* corresponds to two AGRs. Arrows for four examples of optimized replacement codons (*ftsA*, *folC*, *rnpA*, and *rpsJ*) show that deviations in RBS strength and/or mRNA structure are reduced. Arrows are omitted for the remaining eight optimized replacement codons to increase readability.

Kinetic growth analysis showed that the doubling time increased from 52.4 (±2.6) min in EcM2.1 (no AGR codons changed) to 67 (±1.5) min in C123a (123 AGR codons changed in essential genes) in lysogeny broth (LB) at 34 °C in a 96-well plate reader (*Materials and Methods*). Notably, fitness varied significantly during construction of the C123 strain (Fig. 2*B*). This variation may be attributed to codon deoptimization (AGR→CGU) and compensatory spontaneous mutations to alleviate fitness defects in a mismatch repair-deficient (*mutS*-) background. Overall the reduced fitness of C123a may be caused by on-target (AGR→CGU) or off-target (spontaneous) mutations that occurred during strain construction. In this way, *mutS* inactivation is simultaneously a useful evolutionary tool and a liability. Final genome sequence analysis revealed that, along with the 123 desired AGR conversions, C123a had 419 spontaneous nonsynonymous mutations not found in the EcM2.1 parental strain (Fig. S6). Of particular interest was the mutation *argU*_G15A, located in the D arm of tRNA$^{\mathrm{Arg}}$ (*argU*), which arose during CoS-MAGE with AGR set 4. We hypothesized that *argU*_G15A compensates for increased CGU demand and decreased AGR demand, but we observed no direct fitness cost associated with reverting this mutation in C123, and *argU*_G15A does not impact aminoacylation efficiency in vitro or aminoacyl-tRNA pools in vivo (Fig. S7 and Dataset S5). Consistent with the findings of Mukai et al. (25) and Baba et al. (31), argW (tRNA$^{\mathrm{Arg}}{}_{\mathrm{CCU}}$; decodes AGG only) was dispensable in C123a because it can be complemented by argU (tRNA$^{\mathrm{Arg}}{}_{\mathrm{UCU}}$; decodes both AGG and AGA). However, argU is the only *E. coli* tRNA that can decode AGA and remains essential in C123a, probably because it is required to translate the AGR codons for the rest of the proteome (23).

To evaluate the genetic stability of C123a after removal of all AGR codons from all the known essential genes, we passaged C123a for 78 d (640 generations) to test whether AGR codons would recur and/or whether spontaneous mutations would improve fitness. After 78 d, no additional AGR codons were detected in a sequenced population (sequencing data are available at https://github.com/churchlab/agr_recoding), and doubling time of isolated clones ranged from 22% faster to 22% slower than C123a ($n = 60$).

To gain more insight into how local RBS strength and mRNA folding impact codon choice, we performed an evolution experiment to examine the competitive fitness of all 64 possible codon substitutions at each of the AGR codons (Dataset S6). Although MAGE is a powerful method for exploring viable genomic modifications in vivo, we were interested in mapping the fitness cost associated with less-optimal codon choices, requiring codon randomization depleted of the parental genotype, which we hypothesized to be at or near the global fitness maximum. To do so, we developed a method called “CRAM” (Crispr-assisted MAGE). First, we designed oligos that changed not only the target AGR codon to NNN but also made several synonymous changes at least 50 nt downstream that would disrupt a 20-bp CRISPR target locus. MAGE was used to replace each AGR with NNN in parallel, and CRISPR/cas9 was used to deplete the population of cells with the parental genotype. This approach allowed exhaustive exploration of the codon space, including the original codon, but without the preponderance of the parental genotype. Following CRAM, the population was passaged 1:100 every 24 h for 6 d and was sampled before each passage using Illumina sequencing (Fig. 5 and Dataset S6).

Sequencing 24 h after CRAM showed that all codons were present, including stop codons (Fig. S8), validating the method as a technique to generate massive diversity in a population. All sequences for further analysis were amplified by PCR with allele-specific primers containing the changed downstream sequence. Subsequent passaging of these populations revealed many gene-specific trends (Fig. 5 and Figs. S8 and S9). Notably, all codons that required troubleshooting (*dnaT*_AGA10, *ftsA*_AGA19, *frr*_AGA16, and *rnpA*_AGG22) converged to their wild-type AGR codon, suggesting that the original codon was globally optimized. For all cases in which an alternate codon replaced the original AGR, we computed the predicted deviation in mRNA folding energy and local RBS strength (as a proxy for ribosome pausing) for these alternative codons and compared these metrics with the evolution of codon distribution at this position over time. We also computed the fraction of sequences that fall within the SRZ inferred from Fig. 4 (*Materials and Methods*). CRAM initially introduced a large diversity of mRNA folding energies and RBS strengths, but these genotypes rapidly converged toward parameters that are similar to the parental AGR values in many cases (overlays in Fig. 5). Codons that strongly disrupted predicted mRNA folding and internal RBS strength near the start of genes were disfavored after several days of growth, suggesting that these metrics can be used to predict optimal codon substitutions *in silico*. In contrast, nonessential control genes *bcsB* and *chpS* did not converge toward codons that conserved RNA structure or RBS strength, supporting the conclusion that the observed conservation in RNA secondary structure and RBS strength is biologically relevant for essential genes. Interestingly, *tilS_AGA19* was less sensitive to this effect, suggesting that codon choice at that particular position is not under selection. Additionally, the average internal RBS strength for the *ispG* populations converged toward the parental AGR values, but mRNA folding energy averages did not, suggesting that this position in the gene may be more sensitive to RBS disruption than to mRNA folding. Gene *lptF* followed the opposite trend.

Interestingly, several genes (*lptF*, *ispG*, *tilS*, *gyrA*, and *rimN*) preferred codons that changed the amino acid identity from Arg to Pro, Lys, or Glu, suggesting that noncoding functions trump amino acid identity at these positions. Importantly, all successful codon substitutions in essential genes fell within the SRZ (Fig. 6), validating our heuristics based on an unbiased test of all 64 codons. Meanwhile nonessential control gene *chpS* exhibited less dependence on the SRZ.

Discussion

These observations indicate that, although global codon bias may be affected by tRNA availability (6, 42–44), codon choice at a

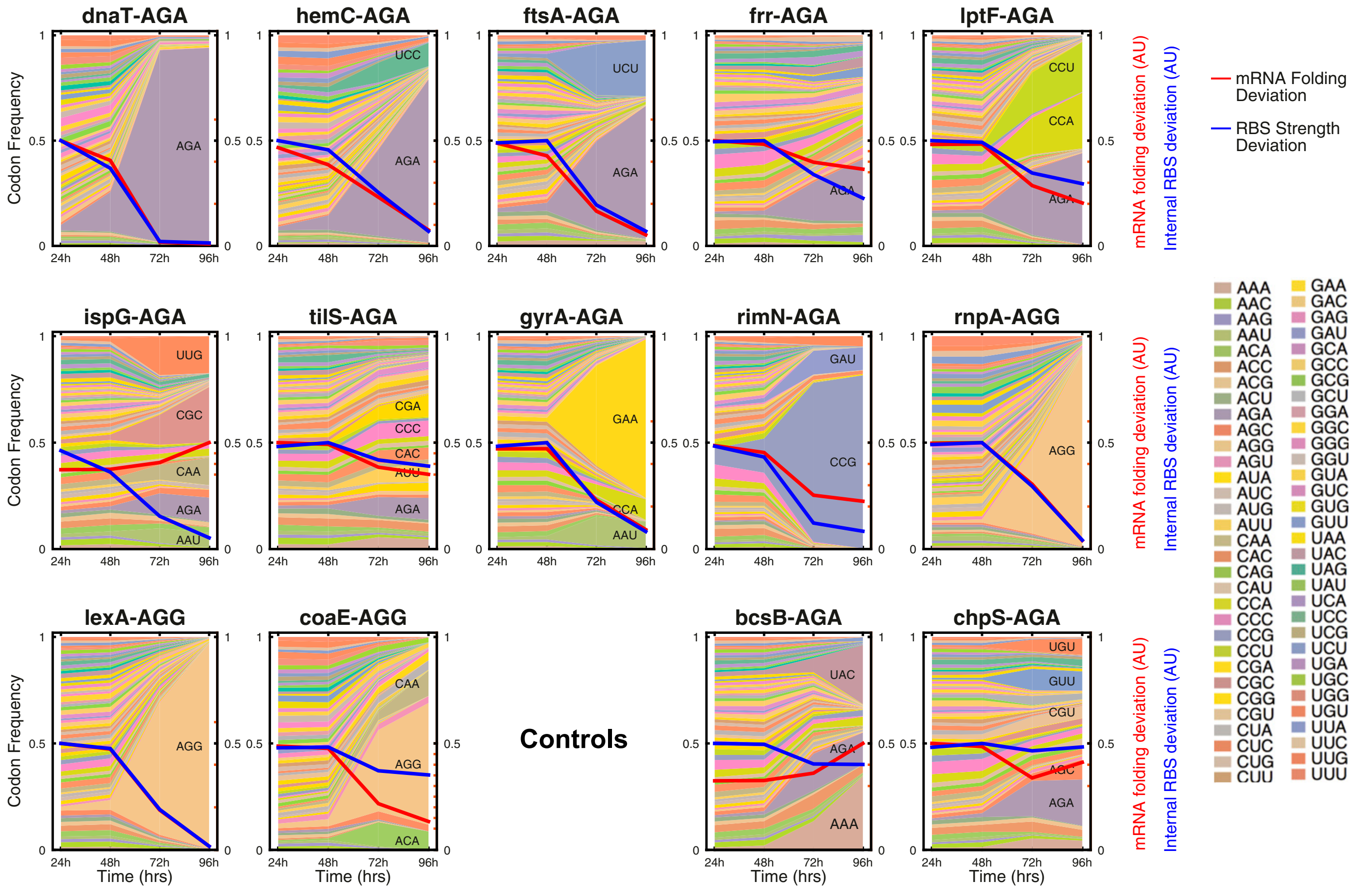

Fig. 5. Codon preference of 14 N-terminal AGR codons. CRAM was used to explore codon preference for several AGR codons located within the first 10 codons of their CDS. Briefly, MAGE was used to diversify a population by randomizing the AGR of interest; then CRISPR/Cas9 was used to deplete the parental (unmodified) population, allowing exhaustive exploration of all 64 codons at a position of interest. Thereafter codon abundance was monitored over time by serially passaging the population of cells and sequencing using an Illumina MiSeq. The left *y* axis (codon frequency) indicates relative abundance of a particular codon (stacked area plot). The right *y* axis indicates the combined deviations in mRNA folding structure (red line) and internal RBS strength (blue line) in arbitrary units (AU) normalized to 0.5 at the initial time point. Zero indicates no deviation from wild type. The horizontal axis indicates the experimental time point (in hours) at which a particular reading of the population diversity was obtained. The genes *bcsB* and *chpS* are nonessential in our strains and thus serve as controls for AGR codons that are not under essential gene pressure.

given position may be defined by at least three parameters: (*i*) amino acid sequence; (*ii*) mRNA structure near the start codon and RBS; and (*iii*) RBS-mediated pausing. In some cases, a subset of these parameters may not be under selection, resulting in an evolved sequence that converges for only a subset of the metrics. In other cases, all metrics may be important, but the primary nucleic acid sequence might not have the flexibility to accommodate all of them equally, resulting in codon substitutions that impair cellular fitness.

These rules were used to generate a draft genome *in silico* with all AGR codons replaced genome-wide, reducing by almost fourfold the number of predicted design flaws (e.g., synonymous codons with metrics outside of the SRZ) as compared with the naive replacement strategy (*Materials and Methods*, Fig. 7, Fig. S10, and Dataset S7). Furthermore, predicting recalcitrant codons provides hypotheses that can be tested rapidly in vivo using MAGE. Successful replacement sequences then can be implemented together in a redesigned genome. Encouragingly, because all newly predicted design flaws occur in nonessential genes, they would be less likely to impact fitness unless (*i*) despite the "nonessential" annotation, the gene is actually essential or quasi-essential (i.e., inactivation would impair growth) or (*ii*) the codon in a nonessential gene impacts the expression of a neighboring essential gene (e.g., impacts an RBS motif or RNA structure). Although incorrect genome annotations can only be addressed empirically (as demonstrated with gene *dedD*), further analysis reveals that AGR codons in nonessential genes should rarely impact annotated essential genes. In *E. coli* MG1655, only three AGR codons in nonessential genes overlap with the initial mRNA and RBS motifs of essential genes, and at least one synonymous CGN codon is predicted to obey the SRZ for all three cases. Furthermore, even if all synonymous mutations were to disobey the SRZ, because disruption of nonessential gene function should not compromise viability, it is expected that nonsynonymous mutations in nonessential genes would be viable as long as they conserve crucial motifs impacting expression of the essential gene. Importantly, we confirmed by MAGE that AGR→CGU codon replacement was possible in two of these three cases and that an alternative synonymous solution could be found in the remaining case (*Materials and Methods*).

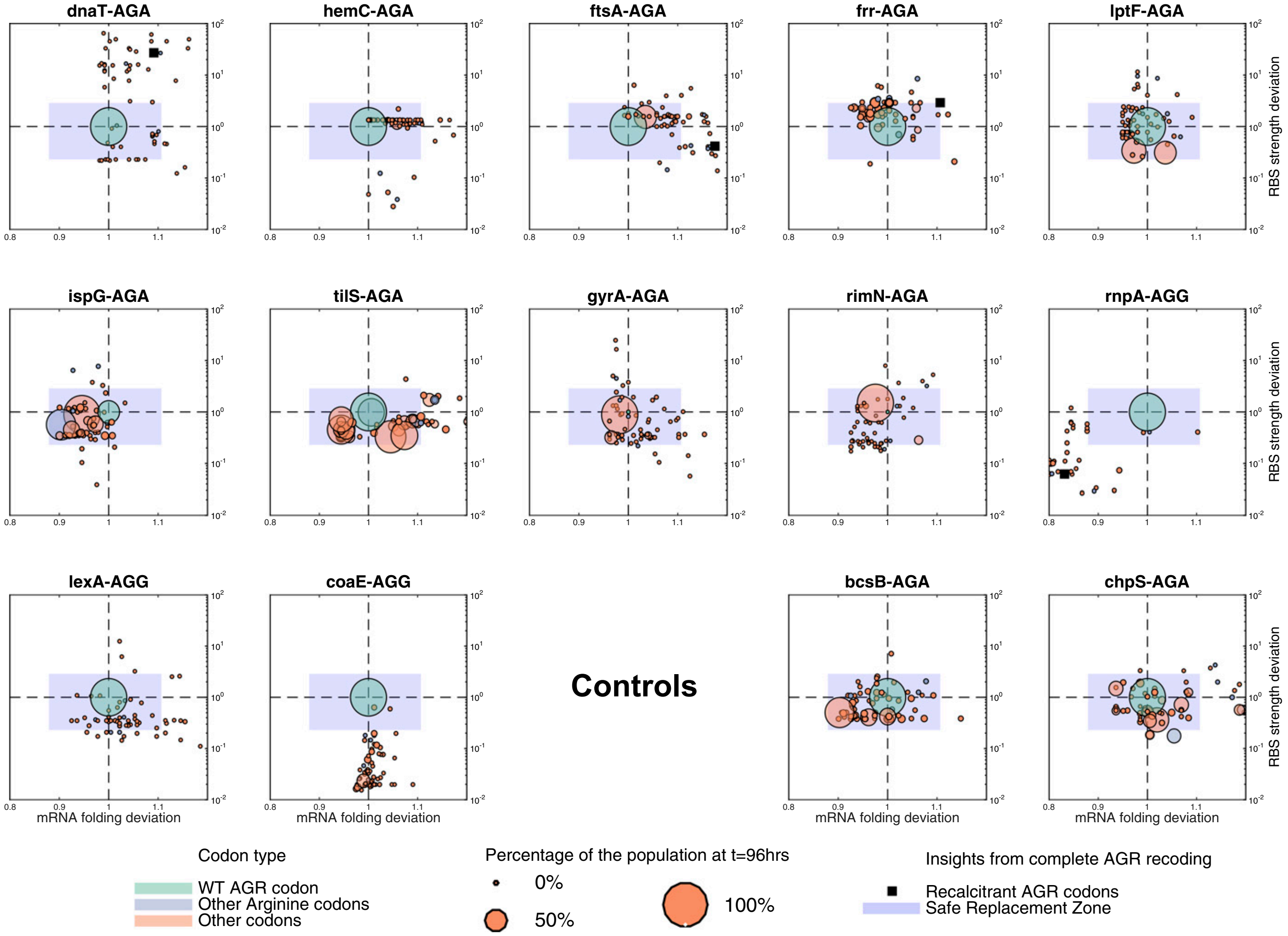

Fig. 6. RBS strength and mRNA structure predict codon preference of 14 N-terminal codon substitutions. Scatter plots show the results of the CRAM experiment (Fig. 5). Each panel represents a different gene. The *y* axis represents RBS strength deviation [calculated with the Salis RBS calculator; (53)], and the *x* axis shows deviations in mRNA folding energy [calculated at 37 °C by the UNAFold calculator (40)]. Codon abundance at the intermediate time point (t = 72 h, chosen to show maximal diversity after selection) is represented by the dot size. Green dots represent the wild-type codon. Blue dots represent synonymous AGR codons. Orange dots represent the remaining 58 nonsynonymous codons, which may introduce nonviable amino acid substitutions. Black squares represent unsuccessful AGR→CGU conversions observed in the genome-wide recoding effort (Fig. 1 and Table 1). The SRZ (blue-shaded region) is the empirically defined range of mRNA folding and RBS strength deviations, based on the successful AGR→CGU replacement mutations observed in this study (Fig. 3). The genes *bcsB* and *chpS* are nonessential in our strains and thus serve as controls for AGR codons that are not under essential gene pressure.

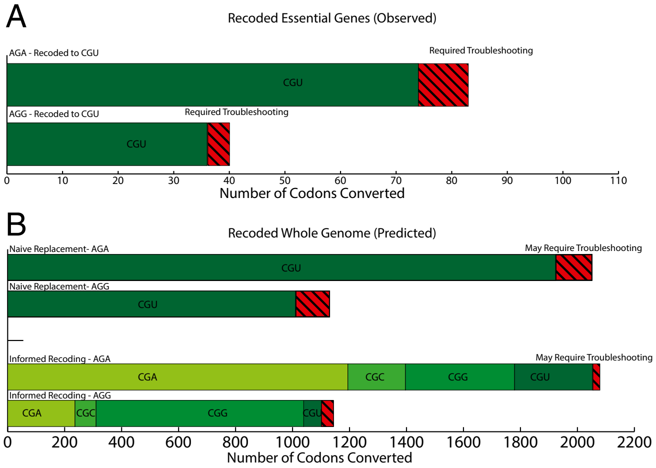

Fig. 7. Predicting optimal replacements for AGR codons reduces the number of codons that are predicted to require troubleshooting. (*A*) Empirical data from the construction of C123. One hundred ten AGR codons were successfully recoded to CGU (green), and 13 recalcitrant AGR codons required troubleshooting (red, striped). (*B*) Predicted recalcitrant codons (codons for which no CGN alternatives fall within the SRZ in Fig. 4) for replacing all instances of the AGR codons genome-wide. The reference genome used for this analysis had insertion elements and prophages removed (54) to reduce total nucleotides synthesized and to increase genome stability, leaving 3,222 AGR codons to be replaced (*Materials and Methods*). Our analysis predicts that replacing all instances of AGR with CGU would have resulted in 229 failed conversions (Naive Replacement, red striped). However, implementing the rules from this work (Informed Replacement) to identify the best CGN alternative reduces the predicted failure rate from 7.1% (229/3,222), to 2.0% (64/3,222) AGR, of which only a small subset will have a direct impact on fitness because the rest are located in nonessential genes. In such cases, MAGE with degenerate oligos could be used to empirically identify replacement codons as we have demonstrated herein. Each specific synonymous CGN is identified by a unique shade of green and is labeled.

Comprehensively removing all instances of AGR codons from all *E. coli* essential genes revealed 13 design flaws that could be explained by a disruption in coding DNA sequence, RBS-mediated translation initiation, RBS-mediated translation pausing, or mRNA structure. Although the importance of each factor has been reported, our work systematically explores the extent to which and the frequency at which they impact genome function. Furthermore, our work establishes quantitative guidelines to reduce the chance of designing nonviable genomes. Although additional factors undoubtedly impact genome function, the fact that these guidelines captured all instances of failed synonymous codon replacements (Fig. 4) suggests that our genome design guidelines provide a strong first approximation of acceptable modifications to the primary sequence of viable genomes. These design rules coupled with inexpensive DNA synthesis will facilitate the construction of radically redesigned genomes exhibiting useful properties such as biocontainment, virus resistance, and expanded amino acid repertoires (45).

Materials and Methods

Strains and Culture Methods. The strains used in this work were derived from EcM2.1 (*E. coli* MG1655 *mutS_mut dnaG_Q576A exoX_mut xonA_mut xseA_mut 1255700::tolQRA* Δ(*ybhB-bioAB*)::[λcI857 N(*cro-ea59*)::*tetR-bla*]) (33). Liquid culture medium consisted of the Lennox formulation of lysogeny broth (LB^L) [1% (wt/vol) bacto tryptone, 0.5% (wt/vol) yeast extract, 0.5% (wt/vol) sodium chloride] (46) with appropriate selective agents: carbenicillin (50 μg/mL) and SDS [0.005% (wt/vol)]. For *tolC* counterselections, colicin E1 (colE1) was used at a 1:100 dilution from an in-house purification (47) that measured 14.4 μg protein/μL (22, 36), and vancomycin was used at 64 μg/mL. Solid culture medium consisted of LB^L autoclaved with 1.5% (wt/vol) Bacto Agar (Fisher), containing the same concentrations of antibiotics as necessary. ColE1 agar plates were generated as described previously (33). Doubling times were determined on a BioTek Eon Microplate reader with orbital shaking at 365 cycles/min at 34 °C overnight and were analyzed using a Matlab script available on GitHub (https://github.com/churchlab/agr_recoding).

Oligonucleotides, PCR, and Isothermal Assembly. A complete table of MAGE oligonucleotides and PCR primers can be found in Dataset S1.

PCR products used in recombination or for Sanger sequencing were amplified with Kapa 2G Fast polymerase according to the manufacturer's standard protocols. Multiplex allele-specific PCR (mascPCR) was used for multiplexed genotyping of AGR-replacement events using the KAPA2G Fast Multiplex PCR Kit, according to previous methods (22, 48). Sanger-sequencing reactions were carried out through a third party (GENEWIZ). CRAM plasmids were assembled from plasmid backbones linearized using PCR (49), and CRISPR/photospacer adjacent motif (PAM) sequences were obtained in Gblocks from Integrated DNA Technologies, using isothermal assembly at 50 °C for 60 min (50).

Lambda Red Recombinations, MAGE, and CoS-MAGE. Lambda Red recombineering, MAGE, and CoS-MAGE were carried out as described previously (33, 51). In singleplex recombinations, the MAGE oligo was used at 1 μM; the coselection oligo was 0.2 μM, and the total oligo pool was 5 μM in multiplex recombinations (7–14 oligos). When double-stranded PCR products were recombined (e.g., tolC insertion), 100 ng of double-stranded PCR product was used. Because we used CoS-MAGE with *tolC* selection to replace target AGR codons, each recombination was paired with a control recombined with water only to monitor *tolC* selection performance. The standard CoS-MAGE protocol for each oligo set was to insert *tolC*, inactivate *tolC*, reactivate *tolC*, and delete *tolC*. mascPCR screening was performed at the *tolC* insertion, inactivation, and deletion steps. All Lambda Red recombinations were followed by a recovery in 3 mL LB^L followed by an SDS selection (*tolC* insertion, *tolC* activation) or ColE1 counterselection (*tolC* inactivation, *tolC* deletion) that was carried out as previously described (33).

General AGR Replacement Strategy. AGR codons in essential genes were found by cross-referencing essential gene annotation according to two complementary resources (31, 52) to find the shared set (107 coding regions), which contained 123 unique AGR codons (82 AGA, 41 AGG). We used optMAGE (35, 51) to design 90-mer oligos (targeting the lagging strand of the replication fork) that convert each AGR to CGU (Datasets S1 and S8). We reduced the total number of AGR replacement oligos to 119 by designing oligos to encode multiple edits where possible, maintaining at least 20 bp of homology on the 5′ and 3′ ends of the oligo. The oligos then were pooled based on chromosomal position into 12 MAGE oligo sets of varying complexity (minimum: 7, maximum: 14) such that a single marker (*tolC*) could be inserted at most 564,622 bp upstream relative to replication direction for all targets within a given set. We then identified *tolC* insertion sites for each of the 12 pools either as intergenic regions or nonessential genes that met the distance criteria for a given pool. See Table 1 for descriptors for each of the 12 oligo pools.

Troubleshooting Strategy. A recalcitrant AGR was defined as one that was not converted to CGU in one of at least 96 clones picked after the third step of the conversion process. The recalcitrant AGR codon then was triaged for troubleshooting (Fig. S1) in the parental strain (EcM2.1). First, the sequence context of the codon was examined for design errors or potential issues, such as misannotation or a disrupted RBS for an overlapping gene. In most cases, corrected oligos could be easily designed and tested. If no such obvious redesign was possible, we attempted to replace AGR with CGN mutations. If attempting to replace AGR with CGN failed to give recombinants, we tested compensatory, synonymous mutations in a 3-aa window around the recalcitrant AGR. If needed, we finally relaxed synonymous stringency by recombining with oligos encoding AGR-to-NNN mutations.

After each step in the troubleshooting workflow, we screened 96 clones from two successive CoS-MAGE recombinations using allele-specific PCR with primers that hybridize to the wild-type genotype. Sequences that failed to yield a wild-type amplicon were Sanger-sequenced to confirm conversion. We also measured doubling time of all clones in LB^L to pair sequencing data with fitness data and chose the recombined clone with the shortest doubling time. Doubling time was determined by obtaining a growth curve on a BioTek plate reader (either an Eon or H1) and was analyzed using web-based open-source genome resequencing software available on GitHub at https://github.com/churchlab/millstone. This genotype then was implemented in the complete strain at the end of strain construction using MAGE and was confirmed by mascPCR screening.

AGR Codons in Nonessential Genes with Impact on Essential Genes. In *E. coli* MG1655, only three AGR codons in nonessential genes overlap with the initial mRNA and RBS motifs of essential genes, and at least one synonymous CGN codon is predicted to obey the SRZ for all three cases. As in the troubleshooting pipeline, we attempted to replace AGR with CGT mutations using MAGE. After four cycles of MAGE, cells were plated, and 96 clones were screened. Synonymous codon replacement was possible for genes *rffT* and *mraW* but not for gene *yidD*. We then relaxed synonymous stringency by recombining with oligos encoding AGR-to-NNN mutations for gene *yidD* and found multiple alternative solutions, including CGA, UGA, GUG, GCG, and TAA. Importantly, the synonymous CGA alternative solutions were less disruptive than CGU to RBS strength and mRNA folding (Dataset S7), further confirming our rules as useful guidelines.

mRNA Folding and RBS Strength Computations. A custom Python pipeline (available at https://github.com/churchlab/agr_recoding) was used to compute mRNA folding and RBS strength value for each sequence. mRNA folding was based on the UNAFold calculator (40) and RBS strength on the Salis calculator (53). The parameters for mRNA folding are the temperature (37 °C) and the window used, which was an average between −30 to +100 nt and −15 to +100 nt around the start site of the gene and was based on ref. 12. The only parameter for RBS strength is the distance between the RBS and the promoter, and we averaged between 9 and 10 nt after the codon of interest based on Li et al. (20). Data visualization was performed through a custom Matlab code.

For *in silico* predictions on the entire genome, all 3,222 AGR in nonphage genes were analyzed using this custom pipeline; data are presented in Dataset S7. Phage genes were not analyzed to reduce the complexity of the genome, inspired by other reduced genome efforts (54).

Whole-Genome Sequencing of Strains Lacking AGR Codons in Their Essential Genes. Sheared genomic DNA was obtained by shearing 130 μL of purified genomic DNA in a Covaris E210 ultrasonicator. Whole-genome library preparation was carried out as previously described (55). Briefly, 130 μL of purified genomic DNA was sheared overnight in a Covaris E210 with the following protocol: duty cycle 10%, intensity 5, 200 cycles per burst, time 780 s per sample. The samples were assayed for shearing on an agarose gel, and, if the distribution was acceptable (peak distribution ∼400 nt), the samples were size-selected by solid-phase reverse immobilization (SPRI)/reverse-SPRI purification as described in ref. 55. The fragments then were blunted, and p5/p7 adaptors were ligated, followed by fill-in and gap repair (New England Biolabs). Then each sample was quantified by quantitative PCR (qPCR) using SYBR green and Kapa Hifi. The results were used to determine how many cycles to amplify the resulting library for barcoding using P5-sol and P7-sol primers. The resulting individual libraries were quantified by NanoDrop (Thermo Scientific) and pooled. The resulting library was quantified by qPCR and an Agilent TapeStation, and run on MiSeq 2 × 150. Data were analyzed to confirm AGR conversions and to identify off-target mutations using Millstone, a web-based open-source genome resequencing tool.

Sequences are available online at https://github.com/churchlab/agr_recoding.

NNN-Sequencing and CRISPR. CRISPR/Cas9 was used to deplete the wild-type parental genotype by selectively cutting chromosomes at unmodified target sites next to the desired AGR codon changes. Candidate sites were determined using the built-in target site finder in Geneious proximally close to the AGR codon being targeted. Sites were chosen if they were less than 50 bp upstream of the AGR codon and could be disrupted with synonymous changes. If multiple sites fulfilled these criteria, the site with the lowest level of sequence similarity to other portions of the genome was chosen. Oligos of a length of ∼130 bp were designed for all 14 genes with an AGR codon in the first 30 nt after the translation start site. Those oligos incorporated both an NNN random codon at the AGR position and multiple (up to six) synonymous changes in a CRISPR target site at least 50 nt downstream of an AGR codon. This change modifies the AGR locus and simultaneously disrupts the CRISPR target site, ensuring randomization of the locus after the parental genotype is deleted.

Specifically, we constructed a plasmid containing the SpCas9 protein gene [plasmid details: DS-SPcas (Addgene plasmid 48645): cloDF13 origin, specR, proC promoter, SPcas9, unused tracrRNA (with native promoter and terminator), J23100 promoter, one repeat (added to facilitate cloning in a spacer onto the same plasmid)]. We also constructed 14 plasmids containing the guide RNA directed toward the unmodified sequences (Plasmid details: PM-! T4Y: p15a origin, chlor^R, J23100 promoter, spacer targeting T4, one repeat).

For each of 24 genes, five cycles of MAGE were performed with the specific mutagenesis oligo at a concentration of 1 μM. CRISPR repeat-spacer plasmids carrying guides designed to target the chosen sites were electroporated into each diversified pool after the last recombineering cycle. After 1 h of recovery, both the SpCas9 and repeat-spacer plasmids were selected for and passaged in three parallel lineages for each of the 24 AGR codons for 144 h. After 2 h of selection, and at every 24-h interval, samples were taken, and the cells were diluted 1/100 in selective medium.

Each randomized population was amplified using PCR primers allowing specific amplification of strains incorporating the CRISPR-site modifications. The resulting triplicate libraries for each AGR codon then were pooled and barcoded with P5-sol and P7-sol primers and run were on a MiSeq 1 × 50. Data were analyzed using custom Matlab code available on https://github.com/churchlab/agr_recoding.

For each gene and each data point, reads were aligned to the reference genome, and frequencies of each codon were computed. In Fig. 5, the mRNA structure deviation (red line) and RBS strength deviation (blue line) in arbitrary units were computed as the product of the frequencies and the corresponding deviation for each codon.

ACKNOWLEDGMENTS. Stephanie Yaung generously provided CRISPR plasmids and technical support. This work was supported by US Department of Energy Grant DE-FG02-02ER63445; by US Defense Advanced Research Projects Agency Grant N66001-12-C-4211 (to F.J.I. and D.S.); by National Institute of General Medical Sciences Grant M22854-42 (to D.S.); by US Department of Defense National Defense Science and Engineering Graduate Fellowships (M.J.L. and G.K.); by a US National Science Foundation Graduate Research Fellowship (D.B.G.); by the Lynch Foundation (M.L.); by an Amazon Web Services in Education Grant Award (G.K.); and by the Arnold & Mabel Beckman Foundation and DuPont, Inc. (F.J.I.). Funding for open access was provided through US Department of Energy Grant DE-FG02-02ER63445.

1. Crick FH (1963) On the genetic code. *Science* 139(3554):461–464.
2. Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 267(5608):275–276.
3. Newton R, Wernisch L (2014) A meta-analysis of multiple matched copy number and transcriptomics data sets for inferring gene regulatory relationships. *PLoS One* 9(8): e105522.
4. dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: A test for translational selection. *Nucleic Acids Res* 32(17):5036–5044.
5. Hershberg R, Petrov DA (2008) Selection on codon bias. *Annu Rev Genet* 42:287–299.
6. Plotkin JB, Kudla G (2011) Synonymous but not the same: The causes and consequences of codon bias. *Nat Rev Genet* 12(1):32–42.
7. Chen GT, Inouye M (1994) Role of the AGA/AGG codons, the rarest codons in global gene expression in Escherichia coli. *Genes Dev* 8(21):2641–2652.
8. Chen GF, Inouye M (1990) Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within the first 25 codons of the Escherichia coli genes. *Nucleic Acids Res* 18(6):1465–1473.
9. Kane JF (1995) Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. *Curr Opin Biotechnol* 6(5):494–500.
10. Sharp PM, Li WH (1987) The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15(3):1281–1295.
11. Sharp PM, Stenico M, Peden JF, Lloyd AT (1993) Codon usage: Mutational bias, translational selection, or both? *Biochem Soc Trans* 21(4):835–841.
12. Goodman DB, Church GM, Kosuri S (2013) Causes and effects of N-terminal codon bias in bacterial genes. *Science* 342(6157):475–479.
13. Zhou M, et al. (2013) Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* 495(7439):111–115.
14. Tuller T, et al. (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* 141(2):344–354.
15. Li GW (2015) How do bacteria tune translation efficiency? *Curr Opin Microbiol* 24: 66–71.
16. Hooper SD, Berg OG (2000) Gradients in nucleotide and codon usage along Escherichia coli genes. *Nucleic Acids Res* 28(18):3517–3523.
17. Gingold H, et al. (2014) A dual program for translation regulation in cellular proliferation and differentiation. *Cell* 158(6):1281–1292.
18. Quax TE, Claassens NJ, Söll D, van der Oost J (2015) Codon bias as a means to fine-tune gene expression. *Mol Cell* 59(2):149–161.
19. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science* 324(5924):255–258.
20. Li GW, Oh E, Weissman JS (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484(7395):538–541.
21. Lajoie MJ, et al. (2013) Probing the limits of genetic recoding in essential genes. *Science* 342(6156):361–363.
22. Isaacs FJ, et al. (2011) Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. *Science* 333(6040):348–353.
23. Lajoie MJ, et al. (2013) Genomically recoded organisms expand biological functions. *Science* 342(6156):357–360.
24. Lee BS, et al. (2015) Incorporation of unnatural amino acids in response to the AGG codon. *ACS Chem Biol* 10(7):1648–1653.

25. Mukai T, et al. (2015) Reassignment of a rare sense codon to a non-canonical amino acid in Escherichia coli. *Nucleic Acids Res* 43(16):8111–8122.
26. Zeng Y, Wang W, Liu WR (2014) Towards reassigning the rare AGG codon in Escherichia coli. *ChemBioChem* 15(12):1750–1754.
27. Rosenberg AH, Goldman E, Dunn JJ, Studier FW, Zubay G (1993) Effects of consecutive AGG codons on translation in Escherichia coli, demonstrated with a versatile codon test system. *J Bacteriol* 175(3):716–722.
28. Spanjaard RA, van Duin J (1988) Translation of the sequence AGG-AGG yields 50% ribosomal frameshift. *Proc Natl Acad Sci USA* 85(21):7967–7971.
29. Spanjaard RA, Chen K, Walker JR, van Duin J (1990) Frameshift suppression at tandem AGA and AGG codons by cloned tRNA genes: Assigning a codon to argU tRNA and T4 tRNA(Arg). *Nucleic Acids Res* 18(17):5031–5036.
30. Bonekamp F, Andersen HD, Christensen T, Jensen KF (1985) Codon-defined ribosomal pausing in Escherichia coli detected by using the pyrE attenuator to probe the coupling between transcription and translation. *Nucleic Acids Res* 13(11):4113–4123.
31. Baba T, et al. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. *Mol Syst Biol* 2:2006.
32. Carr PA, et al. (2012) Enhanced multiplex genome engineering through co-operative oligonucleotide co-selection. *Nucleic Acids Res* 40(17):e132.
33. Gregg CJ, et al. (2014) Rational optimization of tolC as a powerful dual selectable marker for genome engineering. *Nucleic Acids Res* 42(7):4779–4790.
34. Yu D, et al. (2000) An efficient recombination system for chromosome engineering in Escherichia coli. *Proc Natl Acad Sci USA* 97(11):5978–5983.
35. Ellis HM, Yu D, DiTizio T, Court DL (2001) High efficiency mutagenesis, repair, and engineering of chromosomal DNA using single-stranded oligonucleotides. *Proc Natl Acad Sci USA* 98(12):6742–6746.
36. Lajoie MJ, Gregg CJ, Mosberg JA, Washington GC, Church GM (2012) Manipulating replisome dynamics to enhance lambda Red-mediated multiplex genome engineering. *Nucleic Acids Res* 40(22):e170.
37. Ohtake K, et al. (2012) Efficient decoding of the UAG triplet as a full-fledged sense codon enhances the growth of a prfA-deficient strain of Escherichia coli. *J Bacteriol* 194(10):2606–2613.
38. Craigen WJ, Cook RG, Tate WP, Caskey CT (1985) Bacterial peptide chain release factors: Conserved primary structure and possible frameshift regulation of release factor 2. *Proc Natl Acad Sci USA* 82(11):3616–3620.
39. Curran JF (1993) Analysis of effects of tRNA:message stability on frameshift frequency at the Escherichia coli RF2 programmed frameshift site. *Nucleic Acids Res* 21(8): 1837–1843.
40. Markham NR, Zuker M (2008) UNAFold: Software for nucleic acid folding and hybridization. *Methods Mol Biol* 453:3–31.
41. Zadeh JN, et al. (2011) NUPACK: Analysis and design of nucleic acid systems. *J Comput Chem* 32(1):170–173.
42. Novoa EM, Ribas de Pouplana L (2012) Speeding with control: Codon usage, tRNAs, and ribosomes. *Trends Genet* 28(11):574–581.
43. Novoa EM, Pavon-Eternod M, Pan T, Ribas de Pouplana L (2012) A role for tRNA modifications in genome structure and codon usage. *Cell* 149(1):202–213.
44. Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* 2(1):13–34.
45. Lajoie MJ, Söll D, Church GM (2016) Overcoming challenges in engineering the genetic code. *J Mol Biol* 428(5 Pt B):1004–1021.
46. Lennox ES (1955) Transduction of linked genetic characters of the host by bacteriophage P1. *Virology* 1(2):190–206.
47. Schwartz SA, Helinski DR (1971) Purification and characterization of colicin E1. *J Biol Chem* 246(20):6318–6327.
48. Mosberg JA, Gregg CJ, Lajoie MJ, Wang HH, Church GM (2012) Improving lambda red genome engineering in Escherichia coli via rational removal of endogenous nucleases. *PLoS One* 7(9):e44638.
49. Yaung SJ, Esvelt KM, Church GM (2014) CRISPR/Cas9-mediated phage resistance is not impeded by the DNA modifications of phage T4. *PLoS One* 9(6):e98811.
50. Gibson DG, et al. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6(5):343–345.
51. Wang HH, et al. (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460(7257):894–898.
52. Hashimoto M, et al. (2005) Cell size and nucleoid organization of engineered Escherichia coli cells with a reduced genome. *Mol Microbiol* 55(1):137–149.
53. Salis HM (2011) The ribosome binding site calculator. *Methods Enzymol* 498:19–42.
54. Umenhoffer K, et al. (2010) Reduced evolvability of Escherichia coli MDS42, an IS-less cellular chassis for molecular and synthetic biology applications. *Microb Cell Fact* 9:38.
55. Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res* 22(5):939–946.