Genome Biology

METHOD Open Access

Optimizing complex phenotypes through model-guided multiplex genome engineering

Gleb Kuznetsov[1,2,3†], Daniel B. Goodman[1,2†], Gabriel T. Filsinger[1,2,4†], Matthieu Landon[1,4,5], Nadin Rohland[1], John Aach[1], Marc J. Lajoie[1,2*] and George M. Church[1,2*]

Abstract

We present a method for identifying genomic modifications that optimize a complex phenotype through multiplex genome engineering and predictive modeling. We apply our method to identify six single nucleotide mutations that recover 59% of the fitness defect exhibited by the 63-codon *E. coli* strain C321.ΔA. By introducing targeted combinations of changes in multiplex we generate rich genotypic and phenotypic diversity and characterize clones using whole-genome sequencing and doubling time measurements. Regularized multivariate linear regression accurately quantifies individual allelic effects and overcomes bias from hitchhiking mutations and context-dependence of genome editing efficiency that would confound other strategies.

Keywords: Genome engineering, Predictive modeling, Synthetic organisms

Background

Genome editing and DNA synthesis technologies are enabling the construction of engineered organisms with synthetic metabolic pathways [1], reduced and refactored genomes [2–5], and expanded genetic codes [6, 7]. However, genome-scale engineering can come at the cost of reduced fitness or suboptimal traits [2, 7] caused by design flaws that fail to preserve critical biological features [7, 8], synthesis errors, or collateral mutations acquired during strain construction [6]. It remains challenging to identify alleles that contribute to these complex phenotypes and prohibitive to test them individually. Laboratory evolution has traditionally been used to improve desired phenotypes and navigate genetic landscapes [9]; however, this process relies on mutations that accumulate across the genome and may disrupt synthetic designs or traits not maintained under selection. In contrast, targeted genome engineering can alter the genome at chosen loci and can be used to target many locations simultaneously [10]. Multiplexed editing creates a large pool of combinatorial genomic changes than can be screened or selected to find high-performing genomic designs. However, as the number of targeted loci considered increases, it becomes difficult to interpret the significance of individual changes. There remains a need for a method to rapidly identify subsets of beneficial alleles from a large list of candidates in order to optimize large-scale genome engineering efforts.

Leveraging recent improvements in the cost and speed of microbial whole-genome sequencing (WGS), we present a method for identifying precise genomic changes that optimize complex phenotypes, combining multiplex genome engineering, genotyping, and predictive modeling (Fig. 1). Multiple rounds of genome editing are used to generate a population enriched with combinatorial diversity at the targeted loci. Throughout the editing process, clones from the population are subject to WGS and are screened for phenotype. The genotype and phenotype data are used to update a model which predicts the effects of individual alleles. These steps are repeated on a reduced set of candidate alleles informed by the model or on a new set of targets. Finally, the highest impact alleles are rationally introduced into the original organism, minimizing alterations to the organism's original genotype while optimizing the desired phenotype.

* Correspondence: mlajoie@uw.edu; gchurch@genetics.med.harvard.edu
†Equal contributors
[1]Department of Genetics, Harvard Medical School, Boston, MA, USA
Full list of author information is available at the end of the article

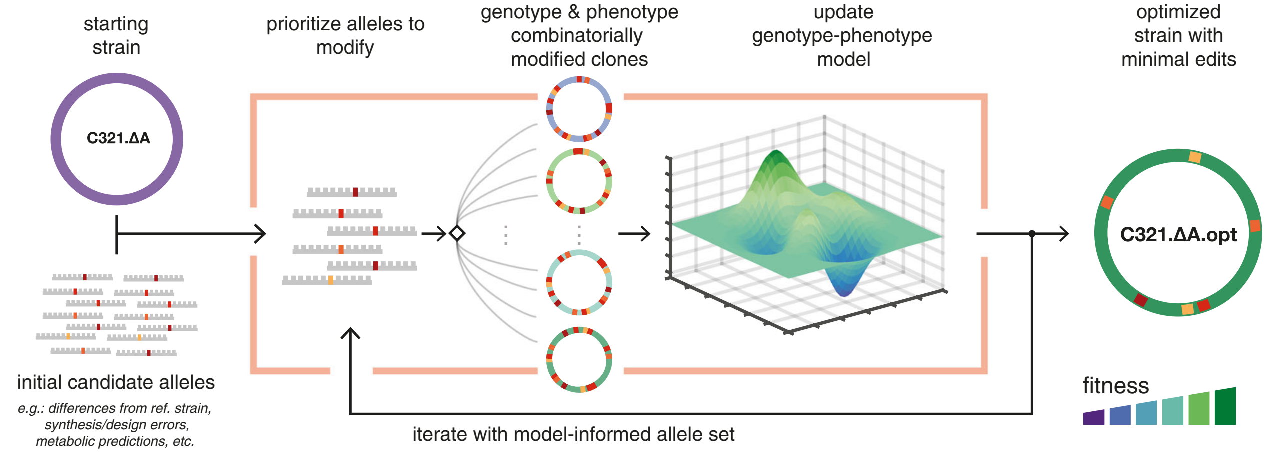
starting strain
C321.ΔA
initial candidate alleles
e.g.: differences from ref. strain, synthesis/design errors, metabolic predictions, etc.
prioritize alleles to modify
genotype & phenotype combinatorially modified clones
update genotype-phenotype model
optimized strain with minimal edits
C321.ΔA.opt
fitness
iterate with model-informed allele set

Fig. 1 Workflow for improving phenotypes through model-guided multiplex genome editing. First, an initial set of target alleles (hundreds to thousands) is chosen for testing based on starting hypotheses. These targets may be designed based on differences from a reference strain, synthesis or design errors, or biophysical modeling. Multiplex genome editing creates a set of modified clones enriched with combinations of the targeted changes. Clones are screened for genotype and phenotype and predictive modeling is used to quantify allele effects. The workflow is repeated to validate and test new alleles. Beneficial alleles are combined to create an optimized genotype

We applied this method to the genomically recoded organism (GRO) C321.ΔA, a strain of *E. coli* engineered for non-standard amino acid (nsAA) incorporation [6]. C321.ΔA was constructed by replacing all 321 annotated UAG stop codons with synonymous UAA codons and deleting UAG-terminating Release Factor 1. Over the course of the construction process, C321.ΔA acquired 355 off-target mutations and developed a 60% greater doubling time relative to its non-recoded parent strain, *E. coli* MG1655. An improved C321.ΔA strain would accelerate the pace of research involving GROs and further enable applications leveraging expanded genetic codes, including biocontainment [11, 12], virus resistance [13], and expanded protein properties [14]. We expected that a subset of the off-target mutations caused a considerable fraction of the fitness defect, providing a starting hypothesis for iterative improvement.

Results

To select an initial set of candidate alleles (Additional file 1: Figure S1), we first used the genome engineering and analysis software *Millstone* [15] to analyze sequencing data from C321.ΔA and to identify all mutations relative to the parental strain MG1655. *Millstone* uses SnpEff [16] to annotate affected genes and predicted severity of each mutation. We further annotated each coding mutation with the growth defect of its associated gene's Keio collection knockout strain after 22 hours in lysogeny broth (LB_22) [17]. Based on this analysis, we identified 127 mutations in proteins and non-coding RNA as the top candidates responsible for fitness impairment. Our candidate alleles included all frameshift and non-synonymous mutations, mutations in non-coding RNA, and synonymous changes in genes with LB_22 < 0.7. We partitioned the targets into three priority categories according to predicted effect (Additional file 2 and Additional file 3).

MAGE introduces combinations of genome edits with approximately 10–20% of cells receiving at least one edit per cycle [10]. To generate a diverse population of mutants enriched for reversions at multiple loci, we performed up to 50 cycles of MAGE in three lineages. The first lineage used a pool of 26 oligonucleotides targeting only the highest category of mutations, the second lineage targeted the top 49 sites, and the third lineage targeted all 127 (Additional file 1: Figure S1).

We sampled a total of 90 clones from multiple time points and lineages during MAGE cycling, including three separate clones of the starting strain. We then performed WGS and measured doubling time for each clone. *Millstone* was used to process sequencing data and to report variants for all 90 samples in parallel [15]. We observed fitness improvement across all three lineages with a diversity of genotypes and fitness phenotypes across the multiple time points (Figs. 2 and 3a, b). Clones selected from the final time point recovered 40–58% (mean 49%) of the fitness defect compared to MG1655 and had 5–15 (mean 10.2) successfully reverted mutations. Of the 127 targeted mutations, 99 were observed in at least one clone, with as many as 19 successful reversions in a clone from the 127-oligo lineage. Additionally, we observed 1329 unique de novo mutations across all clones (although only 135 were called in more than one clone), accumulating at a rate of roughly one per MAGE cycle in each clone (Fig. 2d, e). This elevated mutation rate was caused by defective mismatch

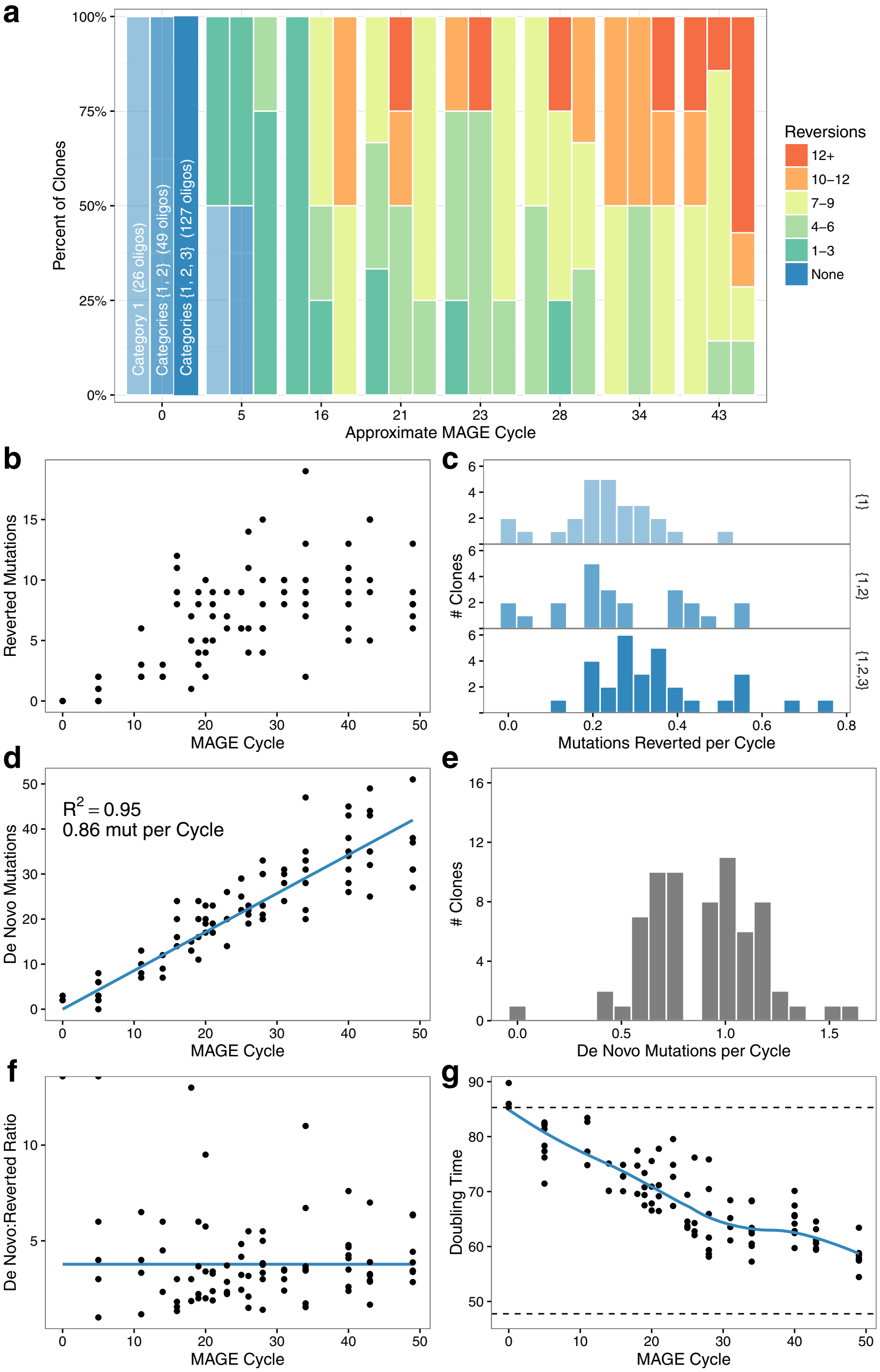

Fig. 2 Mutation dynamics over many cycles of MAGE allele reversion. **a** Increase in combinatorial diversity and reversion count vs. number of MAGE cycles. **b** Number of reversions per clone vs. MAGE cycle. **c** The rate of reversions per MAGE cycle among the different allele categories, showing a higher rate per cycle for cells exposed to all 127 oligos. **d** The number of de novo mutations per clone over successive MAGE cycles. **e** Rate of de novo mutations per MAGE cycle. **f** The average ratio between number of de novo mutations and reverted alleles per MAGE cycle remains constant throughout the experiment. **g** Doubling time (min) improvement per clone from the C321.ΔA starting strain (*top dotted line*) towards the ECNR2 parent strain (*bottom dotted line*). *Blue line* is a LOESS fit

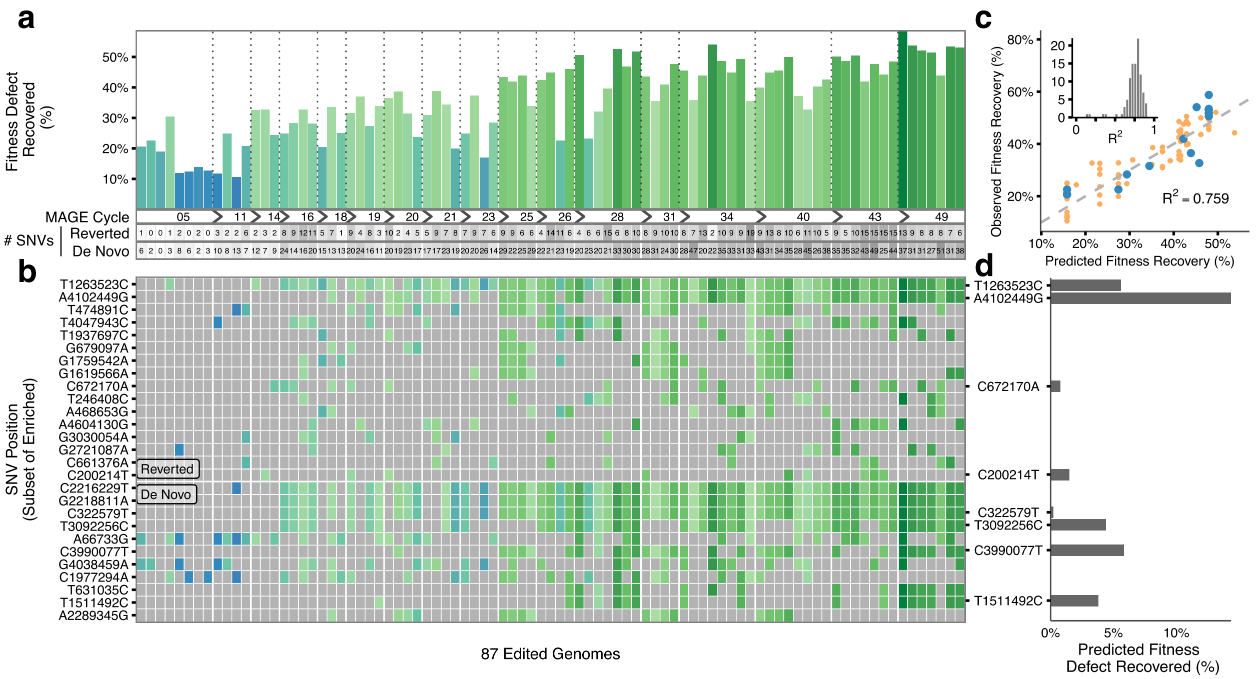

Fig. 3 Genotypic and phenotypic diversity in 87 clones sampled across 50 MAGE cycles enabled model-guided prioritization of top single nucleotide variants (SNVs) for further validation. **a** Percent of C321.ΔA fitness defect recovered across MAGE cycles (shown with bar color and height). The number of SNVs reverted or introduced are shown below. **b** Presence of targeted reversions and de novo mutations in each clone colored according to fitness. A subset of the most enriched mutations is shown, ordered by enrichment (full dataset available in Additional file 10). **c** Example model fit using top eight alleles as features with 15 samples left out as a test set (*blue* points) and used to evaluate R^2. Training points are plotted in *orange*. The inset shows distribution of R^2 values for 100 different simulations with 15 random samples left out to calculate R^2 for each. Example fit was chosen to exemplify a median R^2 value from this distribution. **d** Average model fit coefficients for top eight alleles assigned non-zero values over repeated cross-validated linear regression (see "Methods") indicate their predicted contribution to fitness improvement

repair (Δ*mutS*), which both increases MAGE allele replacement frequency and provides a source of new mutations that could improve fitness.

The combinatorial diversity produced by sampling at regular intervals between consecutive rounds of multiplex genome engineering generates a dataset well suited for analysis by linear regression (Additional file 1: Supplementary Note 3). Initially, we made a simplifying assumption that doubling time is determined by the independent effects of individual alleles and employed a first-order multiplicative model that predicts doubling time based on allele occurrence (see "Methods" and Additional file 1: Supplementary Note 1). As model features, we considered the 99 reversions and 135 de novo mutations that occurred in at least two clones. Multivariate linear regression was used to fit the model, with feature coefficients indicating the predicted effect of the respective allele. We considered several priors in selecting our specific modeling strategy: (1) we expected a small number of alleles to contribute significantly to fitness improvement; (2) the continuous passaging nature of our experiment may allow hitchhiker alleles to become associated with causal alleles. Thus, we chose to use elastic net regularization [18], which adds a weighted combination of L1 and L2 terms to the objective function. To limit overfitting, we performed multiple rounds of k-fold cross-validation (k = 5) and selected alleles that were assigned a non-zero coefficient on average. The analysis of the data obtained over 50 cycles of MAGE identified four targeted reversions and four de novo mutations that had the greatest putative effect on fitness (Fig. 3c, d and Additional file 4).

To validate the eight alleles prioritized in the 50-cycle MAGE experiment, we performed nine cycles of MAGE using a pool of eight oligos (Additional file 4) applied to the starting C321.ΔA strain. We then screened each clone using multiplex allele-specific colony polymerase chain reaction (MASC-PCR) (see "Methods") and measured doubling time (Additional file 1: Figure S2). Modeling revealed strong effects for two reversions (*hemA*-T1263523C and *cpxA*-A4102449G) and one de novo mutation (*cyaA*-C3990077T), along with weaker effects for two additional reversions (*bamA*-C200214T and *leuS*-C672170A). These mutations are discussed in Additional file 1: Supplementary Note 2. A clone with all five of these mutations was isolated and measured to have recovered

51% of the fitness defect exhibited by C321.ΔA. The three remaining de novo mutations did not show evidence of improving fitness despite being highlighted in the initial modeling, illustrating the importance of subsequent validation of model-selected alleles.

To identify mutations that further improved the fitness of C321.ΔA, we extended our search to off-target mutations occurring in regulatory regions using smaller pool sizes. We identified seven non-coding mutations predicted to disrupt gene regulation [8] (see "Methods" and Additional file 5). Applying nine rounds of MAGE followed by linear modeling identified the reversion C49765T, a mutation in the -35 box of the *folA* promoter, which recovers a predicted 27% of the fitness defect (Additional file 1: Figure S3).

To test whether any of the designed UAG-to-UAA mutations caused a fitness defect in the C321 background, we followed the same procedure with 20 previously recoded UAA codons predicted to have a potentially disruptive effect (Additional file 6). We tested reversion back to UAG in a *prfA*$^+$ variant of C321 capable of terminating translation at UAG codons. We observed no evidence of a beneficial fitness effect from any individual UAA-to-UAG reversion.

Finally, we used MAGE to introduce the best six mutations (Additional file 7) into the original C321.ΔA strain (see "Methods"), creating an optimized strain C321.ΔA.opt that restores 59 +/− 11% of the fitness defect in C321.ΔA (Fig. 4a). This rationally designed strain recovered the same amount of fitness as the fastest clones obtained through 50 rounds of MAGE and substantial passaging, which resulted in 6–13 reversions and 31–38 de novo mutations (Fig. 4a). WGS of the final strain confirmed that no UAG codons were reintroduced. Nine additional de novo mutations arose, but these are predicted to have a neutral effect (Additional file 8). We characterized UAG-dependent incorporation of the nsAAs p-acetyl-L-phenylalanine (pAcF) in C321.ΔA.opt using sfGFP variants with 0, 1, and 3 residues replaced by the UAG codon and confirmed that C321.ΔA.opt maintains nsAA-dependent protein expression (Fig. 4b). C321.ΔA.opt has been deposited at AddGene (Bacterial strain #87359).

To address the remaining fitness defect, we first examined potential interactions among the six alleles identified. We characterized the fitness of 359 clones with intermediate genotypes generated during the construction of the final strain (Fig. 4a). We applied linear regression with higher order interaction terms (Fig. 5a) and observed that combinations of mutations tended to produce diminishing returns [19], suggesting that additional beneficial alleles would only contribute marginally to fitness (Fig. 5b). To evaluate the possibility that our modeling procedure did not detect all effects among alleles tested, we performed in silico simulations of a simplified version of our experiment (Additional file 1: Supplementary Note 3) and investigated our ability to detect fitness effect with varying numbers of underlying causal mutations. We found that in the idealized case of no epistasis, we would detect over 90% of total fitness effect given our experimental design parameters (Additional file 1: Figure S4e). A set of relatively weaker mutations may contribute to the remaining fitness defect, although we cannot exclude the possibility that the combination of 321 designed UAG-to-UAA mutations contributes to the global defect as well.

Discussion and conclusion

In summary, we used an iterative strategy of multiplex genome engineering and model-guided feature selection to converge on six alleles that together recover 59% of the fitness defect in C321.ΔA relative to its wild-type ancestor. This method allowed us to quantify the effects of hundreds of individual alleles and then rationally introduce only the minimal set of beneficial genetic changes, reducing unintended effects from additional off-target mutations.

Our approach reveals several problems inherent to simply using enrichment to rank allelic effect. Our data show that alleles enriched over rounds of selection are not necessarily well-correlated with fitness. Allele enrichment may be affected by differences in editing efficiency, competition among beneficial alleles through clonal interference, and genetic drift. Combinatorial targeted editing overcomes these obstacles by allowing the measurement of each allele in many genetic backgrounds, so that linear modeling can quantify its average individual effect.

Further, measuring mutation effects in multiplex makes it experimentally tractable to explore a much larger set of mutations. We observed evidence of positive epistatic interactions between some alleles (Fig. 5a, left), which would be harder to identify through singleplex editing strategies. These findings demonstrate the utility of multiplex genome engineering and predictive modeling for studying epistasis.

A similar model-guided approach could be used to augment other multiplex genome modification techniques, including yeast oligo-mediated genome engineering [19] or multiplex CRISPR/Cas9-based genome engineering in organisms that support homology-directed double-stranded break repair [20, 21]. Biosensors tied to selections or screens [22] can extend this method to optimize biosynthetic pathways in addition to fitness. The rapidly declining cost of multiplex genome sequencing [23] will allow this method to scale to thousands of whole genomes, increasing statistical power and enabling the use of more complex models. While we use column-synthesized oligos in this study, chip-based oligo synthesis enables scaling up the number

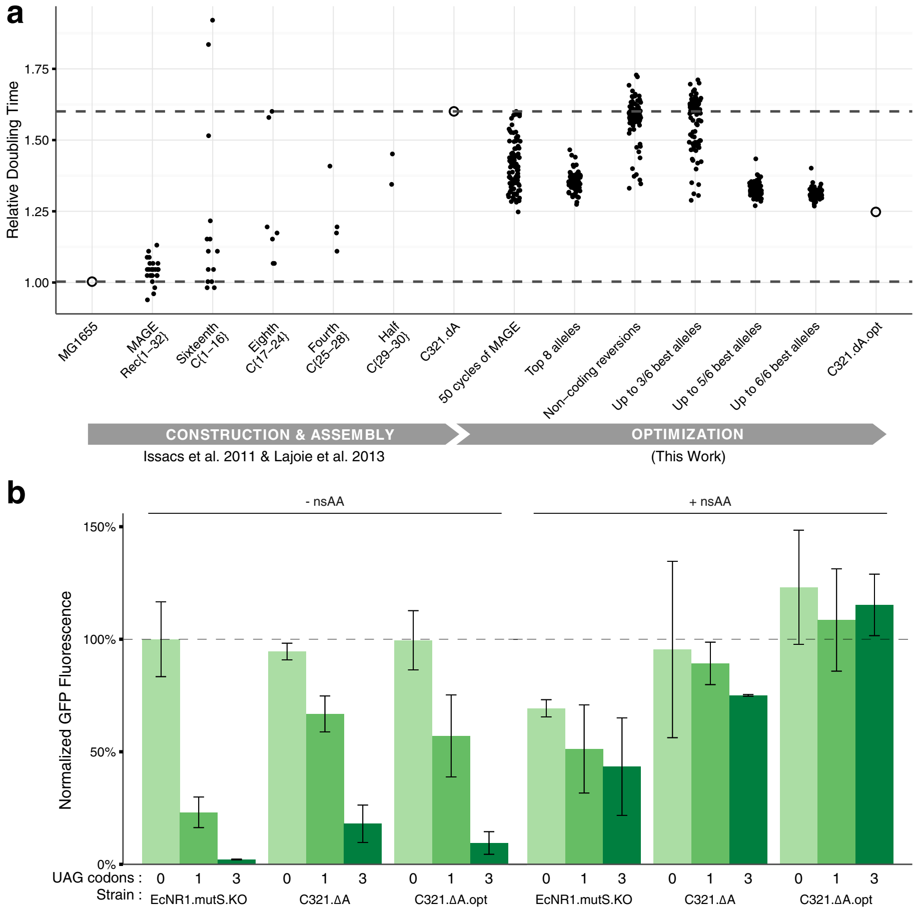

Fig. 4 Construction and characterization of final strain C321.ΔA.opt. **a** Doubling time of clones isolated during construction and optimization of C321.ΔA. Strain C321.ΔA.opt was constructed in seven cycles of MAGE in batches of up to three cycles separated by MASC-PCR screening to pick clones with the maximum number of alleles converted (see "Methods"). The two *dotted horizontal lines* correspond to the relative doubling times for the original GRO and the wild-type strain. **b** Testing nsAA-dependent protein expression using the nsAA p-acetyl-L-phenylalanine (pAcF) in sfGFP variants with 0, 1, or 3 residues replaced with UAG codons. Normalized GFP fluorescence was calculated by taking the ratio of absolute fluorescence to OD600 of cells suspended in phosphate buffered saline (PBS) for each sample and normalizing to the fluorescence ratio of non-recoded strain EcNR1.mutS.KO expressing 0 UAG sfGFP plasmid

of genomic sites targeted, allowing thousands of alleles to be tested simultaneously [24–26]. Our simulations suggest that the predictive power of this method can support larger number of mutations than we tested with a modest increase in genomes sampled (Additional file 1: Figure S4d). Finally, making genomic changes trackable [27–29] for targeted sequencing could further increase the economy, speed, and throughput of this approach.

Efficiently quantifying the effects of many alleles on complex phenotypes is critical not only for tuning synthetic organisms and improving industrially relevant phenotypes, but also for understanding genome architecture. While our method is used here to identify and repair detrimental alleles to improve fitness, it will also enable rapid prototyping of alternative genome designs and interrogation of genomic design constraints. Iteratively measuring and modeling the effects of large numbers of combinatorial genomic changes in parallel is a powerful approach to navigate and understand genotype-phenotype landscapes.

Methods

Media and reagents

All experiments were performed in LB-Lennox (LBL) medium (10 g/L bacto tryptone, 5 g/L sodium chloride, 5 g/L yeast extract) with pH adjusted to 7.45 using 10 M NaOH. LBL agar plates were made from LBL plus 15 g/L Bacto Agar. Selective agents were used at the following concentrations: carbenicillin (50 μg/mL), chloramphenicol (20 μg/mL), gentamycin (5 μg/mL), kanamycin (30 μg/mL), spectinomycin (95 μg/mL), and SDS (0.005% w/v).

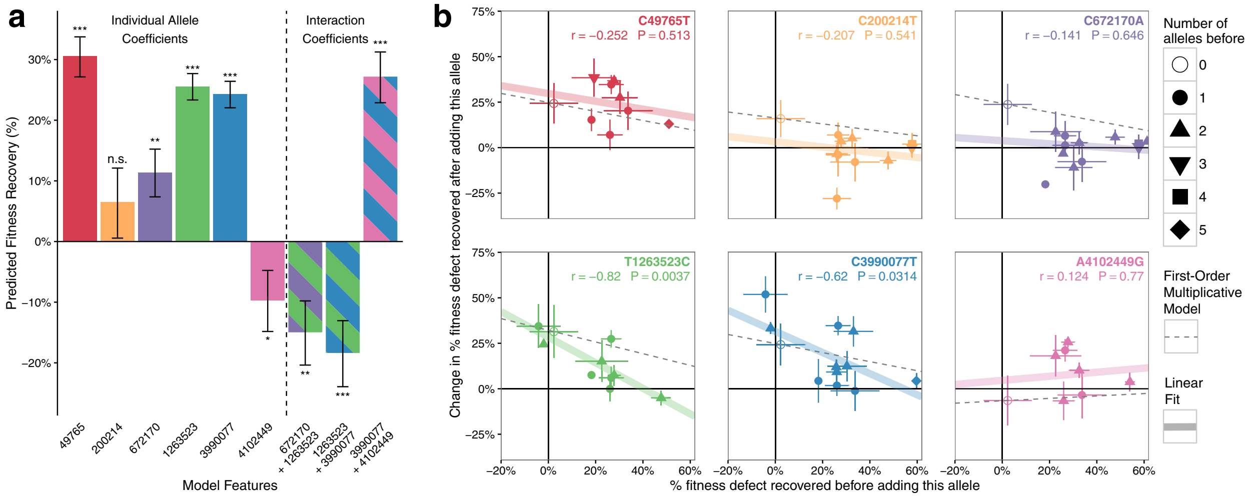

Fig. 5 Interactions among top six alleles show evidence of epistasis. Genotypes and fitness measurements were obtained from 359 intermediate clones generated during the construction of the final strain containing the six best alleles (Additional file 7). Each clone was genotyped using MASC-PCR and doubling time was measured during allele validation experiments and final strain construction. **a** Individual model coefficients for the top six alleles, as well as three significant interaction terms identified during combinatorial construction. These values are from a linear model with interaction terms between each pair of alleles. The *error bars* signify the standard error of the mean of the model coefficients and the significance codes for a non-zero effect size are: *** $p < 0.001$, ** $0.001 \leq p < 0.01$, * $0.01 \leq p < 0.05$, *n.s.* not significant. All three interactions coefficients remain significant after a family-wise error rate (FWER) of $\alpha = 0.05/C(6,2) = 0.003$. **b** Each data point represents the amount of fitness recovered when adding the allele specified to an identical starting genotype background. *Horizontal error bars* correspond to the standard deviation of fitness defect among all clones with this starting genotype. *Vertical error bars* represent the standard deviation of all differences between clones with and without the respective allele. For each plot, the thick *colored line* represents a simple linear fit through the points, corresponding to the *r* and *p* values given in each plot. The *dotted line* corresponds to the predicted fit for a simple multiplicative model of fitness where the allele always recovers a constant percent of the remaining fitness defect regardless of the background. For all alleles except A4102449G (*pink*), adding the allele to C321 showed a recovery of the fitness defect (>0 on the y axis), with the percentage of defect recovered decreasing as other alleles are also reverted, consistent with a first-order multiplicative model. In some cases, the fitness improvement drops more rapidly than predicted by the multiplicative model (i.e. points below the *dotted lines*), suggesting diminishing returns epistasis. This is supported by the negative-coefficient interaction terms in panel (**a**). In the case of A4102449G there appears to be a negative effect with the mutation alone, but an increase in the presence of other alleles, suggesting possible sign epistasis

Strains

The construction and genotype of engineered *E. coli* strain C321.ΔA was previously described in detail [6]. Here, before improving fitness, we constructed strain *C321.ΔA.mutSfix.KO.tolCfix.Δbla:E* by further modifying C321.ΔA to introduce the following changes: (1) the *mutS* gene was reinserted into the C321.ΔA strain in its original locus and MAGE was used to disable the gene by introduction of two internal stop codons and a frameshift; and (2) the carbenicillin-resistance marker *bla* was swapped for gentamicin resistance marker *aacC1* in the lambda red insertion locus. Several control assays were performed in EcNR1.mutS.KO, a non-recoded by MAGE-enabled strain similar to EcNR2 [10]. All genomic positions reported in the manuscript are in the frame of MG1655 K12 (Genbank accession NC_000913.2). The final C321.ΔA.opt strain has been deposited at AddGene (Bacterial strain #87359).

Millstone, software for multiplex genome analysis and engineering

Millstone [15] was used throughout the project to rapidly process WGS data and identify variants in each sample relative to the reference genome, to explore variant data, and to design oligonucleotides for MAGE. The *Millstone* analysis pipeline takes as input raw FASTQ reads for up to hundreds of clones and a reference genome as Genbank or FASTA format. The software then automates alignment of reads to the reference using the Burrows-Wheeler Aligner (BWA-MEM) followed by single nucleotide variant (SNV) calling using Freebayes. *Millstone* performs variant calling in diploid mode, even for bacterial genomes. This helps account for paralogy in the genome and results in mutation calls being reported as “homozygous alternate” (strong wild-type), “heterozygous” (marginal), or wild-type, along with an “alternate fraction” (AF) field that quantifies the fraction of aligned reads at the locus showing the alternate allele. Marginal calls were inspected on a case-by-case basis using *Millstone*’s JBrowse integration to visualize raw read alignments. *Millstone* provides an interface for exploring and comparing variants across samples. After initial exploration and triage in *Millstone*, we exported the variant report from *Millstone* for further analysis and predictive modeling. In follow-up analysis, we determined empirically that $0.1 < AF < 0.7$ indicated a variant call was marginal in our data.

Identifying off-target mutations for reversions

For the 50-cycle MAGE experiment, we considered only mutations occurring in regions annotated as coding for a protein or functional RNA. Using *Millstone* annotations of predicted effect and Keio knock-out collection annotation of essentiality [17], we defined three priority categories according to expected effect on fitness (Additional file 2). A total of 127 targets were allocated to the three categories to be used for the 50-cycle MAGE experiment.

For a separate experiment, off-target mutations in regulatory regions were selected based on the criteria of predicted regulatory disruption of essential genes and several non-essential genes with particularly strong predicted disruption. Regulatory disruption was determined based on calculating change in 5′ messenger RNA (mRNA) folding or ribosome binding site (RBS) motif strength for mutations occurring up to 30 bases upstream of a gene. We calculated mRNA folding and RBS motif disruption as described in [8]. Briefly, the minimum free energy (MFE) of the 5-prime mRNA structure was calculated using Unafold's hybrid-ss-min function [30] ($T = 37$ °C), taking the average MFE between windows of RNA (−30, +100) and (−15, +100) relative to the start codon of the gene. Mutations that caused a change in MFE of the mRNA of over 10% relative to the wild-type context were prioritized for testing. To predict RBS disruption, the Salis RBS Calculator [31] was provided with sequence starting 20 bases upstream of the gene ATG and including the ATG. Mutations that caused a greater than tenfold change in predicted expression were included for testing. Finally, we also considered mutations that overlapped promoters of essential genes based on annotations from RegulonDB [32].

The 20 UAG-reversion targets were chosen when UAGs occurred in essential genes, introduced non-synonymous changes in overlapping genes, or disrupted a predicted regulatory feature as above.

Multiplex automated genome engineering

Single-stranded DNA oligonucleotides for MAGE were designed using *Millstone*'s optMAGE integration (https://github.com/churchlab/optmage). Oligos were designed to be 90 bp long with the mutation located at least 20 bp away from either end. We used the C321.ΔA reference genome (Genbank accession CP006698.1) for oligo design to avoid inadvertently reverting intentional UAG-to-UAA changes. OptMAGE avoids strong secondary structure (< -12 kcal mol − 1) and chooses the sense of the oligo to target the lagging strand of the replication fork [10]. Phosphorothioate bonds were introduced between the first and second and second and third nucleotides at the 5-prime end of each oligo to inhibit exonuclease degradation [10]. All DNA oligonucleotides were purchased with standard purification and desalting from Integrated DNA Technologies and dissolved in dH20.

MAGE was performed as described in [10], with the following specifications: (1) cells were grown at 34 °C between cycles; (2) we noted that C321.ΔA exhibits electroporation resistance so a voltage of 2.2 kV (BioRad GenePulser, 2.2 kV, 200 ohms, 25 μF was used for cuvettes with 1 mm gap) was chosen based on optimization using a lacZ blue-white screen; and (3) total concentration of the DNA oligonucleotide mixture was 5 μM for all electroporations (i.e. the concentration of each oligo was adjusted depending on how many oligos were included in the pool).

The 50-cycle MAGE experiment was carried out in three lineages, with oligo pool sizes of 26, 49, and 127 consisting of oligos from priority categories {1}, {1,2}, and {1,2,3}, respectively (Additional file 2). Note that we originally began with just two pools—the top 26 and all 127 oligos—but after five MAGE cycles the lineage exposed to all 127 oligos was branched to have a separate lineage with only the 49 category {1, 2} oligos in order to obtain more enrichment of the higher priority targets. In order to prevent any population from acquiring permanent resistance to recombination, we toggled the dual-selectable marker tolC at recombinations 23, 31, and 26 for the three lineages, respectively, as described in [32]. Briefly, an oligo introducing an internal stop codon in *tolC* was included in the recombination, and after at least 5 h of recovery, cells were selected in media containing colicin E1, which is toxic in *tolC*$^{+}$*E. coli*. In the subsequent recombination, an oligo restoring *tolC* function was included in the pool after which cells were selected in the presence of 0.005% SDS (w/v).

Validation MAGE experiments composed of ten or fewer oligos were carried out for up to nine MAGE cycles, as we expected adequate diversity based on previous experience with MAGE efficiency.

Whole-genome sequencing

Genomic DNA (gDNA) preparation for WGS of 96 clones (only 87 considered in manuscript because sequencing analysis revealed that nine cultures were polyclonal) was performed as in [33]. Briefly, gDNA was prepared by shearing using a Covaris E210 AFA Ultrasonication machine. Illumina libraries were prepared for pooled sequencing as previously described [34]. Barcoded Illumina adapters were used to barcode each strain in a 96-well plate. All 96 genomes were sequenced together on a single lane of a HiSeq 2500 PE150 (Additional file 9). Alternative inexpensive WGS library preparation methods have since become available [23].

WGS data were processed to identify clonal genotypes in *Millstone* and then exported for further analysis (Additional file 10). Demultiplexed.fastq reads were aligned to the MG1655 reference genome. SNVs were reported with *Millstone*, as described above. During analysis, marginal calls were visually confirmed by examining alignments using *Millstone*'s JBrowse integration.

MASC-PCR

MASC-PCR was used to assess successful reversions in validation experiments of ten or fewer targeted mutations and typically performed for 96 clones in parallel. The protocol was performed as previously described [6]. Briefly, two separate PCRs, each interrogating up to ten positions simultaneously, were performed on each clone to detect whether the C321.ΔA or reverted allele was present at each position. For each position, the two reactions shared a common reverse primer but used distinct forward primers differing in at least one nucleotide at the 3′ end to match the SNV being assayed specifically. Positive and negative controls were included when available to aid in discriminating cases of non-specific amplification.

Measuring fitness

Fitness was determined from kinetic growth (OD600) on a Biotek H-series plate reader. Cells were grown at 34 °C in 150 μL LBL in a flat-bottom 96-well plate at 300 rpm linear shaking. To achieve consistent cell state before reading, clones were picked from agar plates or glycerol, grown overnight to confluence, passaged 1:100 into fresh media, grown again to mid-log (~3 h), and passaged 1:100 again before starting the read. OD measurements were recorded at 5-min intervals until confluence. Doubling times were calculated according to $t_{double} = c * \ln(2)/m$, where c = 5 min per time point and m is the maximum slope of ln(OD600). The maximum slope was determined using a sliding window linear regression through eight contiguous time points (40 min) points rather than between two predetermined OD600 values because not all of the growth curves were the same shape or reached the same max OD600. The script used for analyzing doubling time is available at https://github.com/churchlab/analyze_plate_reader_growth.

Predictive modeling of allele causality

Choosing alleles for subsequent validation was framed as a feature selection problem. We used predictive modeling to prioritize features. Both targeted reversions introduced by MAGE and de novo mutations were considered.

For most analyses, we used a first-order multiplicative allele effect model, where each allele (reversion or de novo mutation) is represented by a single feature and the fitted coefficient corresponding to that feature represents the allele's effect on doubling time. To find coefficient values, we fit a linear model where genotypes (WGS or MASC-PCR) predict the logarithm of doubling time. Alleles corresponding to features with the most negative coefficients were selected for validation in smaller sets. An additive model was also tested and yielded similar results, as previously noted by others [19].

While we anticipated the possibility of epistatic effects among alleles tested, a first-order model of the 50-cycle MAGE experiment already had 239 features (99 reversions + 140 de novos observed at least twice) and 87 samples, so we omitted higher-order interaction terms to avoid overfitting due to model complexity. We discuss implications of this independence assumption and other details of our allele effect modeling strategy in Additional file 1: Supplementary Note 1.

Elastic net regularization [18], which includes both L1 and L2 regularization penalties, was used in model-fitting. L1 regularization enforces sparsity, capturing the assumption that a handful of alleles will explain a majority of the fitness effect. L2 regularization prevents any one of a subset of highly correlated alleles from dominating the effect of those alleles, balancing the tendency of L1 to drop subsets of highly co-occurring alleles.

Accordingly, the elastic net loss function used follows from Zou and Hastie [18]:

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_1 |\beta|_1 + \lambda_2 |\beta|^2$$

where

$$|\beta|_1 = \sum_{j=1}^{p} \left|\beta_j\right|$$

$$|\beta|^2 = \sum_{j=1}^{p} \beta_j^2$$

And the coefficients were estimated according to:

$$\hat{\beta} = argmin_{\beta}(L(\lambda_1, \lambda_2, \beta))$$

Elastic net regression was performed using the ElasticNetCV module from scikit-learn [35]. This module introduces the hyperparameters $alpha = \lambda_1 + \lambda_2$ and $l1_{ratio} = \frac{\lambda_1}{\lambda_1, + \lambda_2}$ and uses k-fold cross validation (k = 5) to identify the best choice of hyperparameters for a given training dataset. We specified the range of l1_ratio to search over as [0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99, 1], which tests with higher resolution near L1-only penalty. This fits our hypothesis that a small number of mutations are responsible for a majority of the fitness effect. For alpha, we followed the default of allowing scikit-learn to search over 100 alpha values automatically computed based on l1_ratio.

To avoid overfitting due to the undersampled nature of the data in the 50-cycle MAGE experiment, we performed 100 repetitions of scikit-learn's cross-validated elastic net regression procedure, and for each repetition, we randomly held out 15 samples that could be used to evaluate the model fit by that iteration. The model coefficient for each allele was then calculated as the weighted-average across all 100 repetitions using the prediction score on the 15 held-out samples as the weighting factor. Only model coefficients with a negative value (some putative fitness improvement) were considered in a second round of 100 repeats of cross-validated elastic net regression, again with 15 samples held out in each repeat to evaluate the model fit. The weighted-average coefficient values over this second set of 100 repetitions were used to determine the top alleles for experimental validation in a nine-cycle MAGE experiment. While this method reproducibly reported the alleles *hemA*-T1263523C, *cpxA*-A4102449G, and *cyaA*-C3990077T, alleles with weaker predicted effects were detected more stochastically, depending on the randomized train-test split, even with 100 repetitions. We expect that sequencing additional clones, as well as further tuning of our modeling method for detecting weak effects may be warranted in future studies.

To evaluate the results of the nine-cycle MAGE validation experiments, we used unregularized multivariate linear regression. With ten or fewer parameters and ~90 clones, only a single iteration of cross-validated regression applied to the full dataset was required to assign predicted effects without requiring the testing of individual alleles.

Elastic net-regularized multivariate regression was compared to univariate linear regression for our data (Additional file 1: Supplementary Note 1, Additional file 11).

Final strain construction

C321.ΔA.opt was constructed by adding the six alleles identified by the optimization workflow (Additional file 7) to C321.ΔA.mutSfix.KO.tolCfix.Δbla:E. A total of seven cycles of MAGE were required, with a MASC-PCR screening step every three cycles to select a clone with the best genotype so far (Fig. 3a), minimizing the total number of cycles required. Three cycles of MAGE were performed using oligos targeting all six alleles. Ninety-six clones were screened by MASC-PCR, and one clone with 3/6 alleles (C49765T, T1263523C, A4102449G) was chosen for the next round of MAGE. Three more rounds of MAGE were performed on top of the clone with 3/6 alleles using only the three remaining oligos. MASC-PCR identified a clone with 5/6 alleles (C49765T, C200214T, C672170A, T1263523C, A4102449G). One more round of MAGE was performed using the remaining oligo and a clone with all six alleles was obtained. Additional off-target mutations acquired during construction as identified by whole genome sequencing of the final clone are listed in Additional file 8.

Characterizing non-standard amino acid incorporation

nsAA incorporation was measured as previously described [6]. 1-UAG-sfGFP, and 3-UAG-sfGFP reporters were produced by PCR mutagenesis from sfGFP (Additional file 1: Supplementary Note 4), and isothermal assembly was used to clone 0-UAG-sfGFP (unmodified sfGFP), 1-UAG-sfGFP, and 3-UAG-sfGFP into the pZE21 vector backbone [36]. We used the pEVOL-pAcF plasmid to incorporate the non-standard amino acid p-acetyl-L-phenylalanine. Reagents were used at the following concentrations: anhydrotetracycline (30 ng/μL), L-arabinose (0.2% w/v), pAcF (1 mM).

Additional files

Additional file 1: Supplementary Figures, Notes, and Table Legends [37–45]. (PDF 1142 kb)

Additional file 2: Table S1. Prioritized coding reversion categories. (XLSX 36 kb)

Additional file 3: Table S2. Table of 127 mutations targeted in the 50-cycle MAGE experiment. (XLSX 50 kb)

Additional file 4: Table S3. Top mutations from MAGE cycling. (XLSX 37 kb)

Additional file 5: Table S4. List of non-coding mutations tested. (XLSX 8 kb)

Additional file 6: Table S5. List of amber reversions tested. (XLSX 9 kb)

Additional file 7: Table S6. List of top six mutations used to create C321.DA.opt. (XLSX 36 kb)

Additional file 8: Table S7. List of additional mutations in C321.DA.opt. (XLSX 39 kb)

Additional file 9: Table S8. Cost analysis. (XLSX 35 kb)

Additional file 10: Table S9. Mutations from whole genome sequencing data for 90 clones from 50-cycle MAGE experiment. (XLSX 12780 kb)

Additional file 11: Table S10. Comparison to univariate model (GWAS). (XLSX 45 kb)

Acknowledgements

We thank members of the Church Lab for comments and Alex Rothberg for helpful suggestions on predictive modeling.

Funding

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE-FG02-02ER63445. GK and MJL were supported by DOD NDSEG Fellowships. DBG and GTF were supported by NSF Graduate Research Fellowships. Computational resources for this work were provided by the AWS Cloud Credits for Research Program.

Availability of data and materials

Analysis and simulation code is available at https://github.com/churchlab/optimizing-complex-phenotypes, released under the MIT license. WGS data for intermediate clones, the final strain C321.ΔA.opt, and the C321.ΔA.opt ancestor C321.ΔA.mutSfix.KO.tolCfix.Δbla:E have been deposited at the NIH Sequence Read Archive (SRA) under BioProject PRJNA382959. C321.ΔA.opt may be requested from AddGene (Bacterial strain #87359).

Authors' contributions

GK, MJL and DBG designed the study. GK, MJL, GTF, DBG and ML designed and performed experiments. NR assisted with whole genome sequencing. GK, DBG, and GTF performed data analysis. GK, GTF, DBG, JA and MJL wrote the manuscript; all authors contributed to editing of the manuscript. GMC supervised the project. All authors read and approved the final manuscript.

Competing interests

GK, DBG, ML, MJL, and GMC are inventors on patent application #62350468 submitted by the President and Fellows of Harvard College, covering methods for designing genomes. GMC serves on the Scientific Advisory Board (SAB) and has a financial interest in ReadCoor, GRO Biosciences and EnEvolv, Inc., that are related to the above referenced grant (DE-FG02-02ER63445). Additionally, MJL has a financial interest in GRO Biosciences. For GMC other potentially relevant financial interests are listed at arep.med.harvard.edu/gmc/tech.html.

Ethics approval and consent to participate

Not applicable.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Genetics, Harvard Medical School, Boston, MA, USA. [2]Wyss Institute for Biologically Inspired Engineering, Harvard Medical School, Boston, MA, USA. [3]Program in Biophysics, Harvard University, Boston, MA, USA. [4]Systems Biology Graduate Program, Harvard Medical School, Boston, MA, USA. [5]Ecole des Mines de Paris, Mines Paristech, Paris, France.

Received: 2 December 2016 Accepted: 21 April 2017
Published online: 25 May 2017

References

1. Wei T, Cheng B-Y, Liu J-Z. Genome engineering Escherichia coli for L-DOPA overproduction from glucose. Sci Rep. 2016;6:30080.
2. Hutchison 3rd CA, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, et al. Design and synthesis of a minimal bacterial genome. Science. 2016;351:aad6253.
3. Pósfai G, Plunkett 3rd G, Fehér T, Frisch D, Keil GM, Umenhoffer K, et al. Emergent properties of reduced-genome Escherichia coli. Science. 2006;312:1044–6.
4. Chan LY, Kosuri S, Endy D. Refactoring bacteriophage T7. Mol Syst Biol. 2005;1:2005.0018.
5. Dymond JS, Richardson SM, Coombes CE, Babatz T, Muller H, Annaluru N, et al. Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. Nature. 2011;477:471–6.
6. Lajoie MJ, Rovner AJ, Goodman DB, Aerni H-R, Haimovich AD, Kuznetsov G, et al. Genomically recoded organisms expand biological functions. Science. 2013;342:357–60.
7. Ostrov N, Landon M, Guell M, Kuznetsov G, Teramoto J, Cervantes N, et al. Design, synthesis, and testing toward a 57-codon genome. Science. 2016; 353:819–22.
8. Napolitano MG, Landon M, Gregg CJ, Lajoie MJ, Govindarajan L, Mosberg JA, et al. Emergent rules for codon choice elucidated by editing rare arginine codons in Escherichia coli. Proc Natl Acad Sci U S A. 2016;113: E5588–97.
9. Dragosits M, Mattanovich D. Adaptive laboratory evolution – principles and applications for biotechnology. Microb Cell Fact. 2013;12:64.
10. Wang HH, Isaacs FJ, Carr PA, Sun ZZ, George X, Forest CR, et al. Programming cells by multiplex genome engineering and accelerated evolution. Nature. 2009;460:894–8.
11. Mandell DJ, Lajoie MJ, Mee MT, Takeuchi R, Kuznetsov G, Norville JE, et al. Biocontainment of genetically modified organisms by synthetic protein design. Nature. 2015;518:55–60.
12. Rovner AJ, et al. "Recoded organisms engineered to depend on synthetic amino acids." Nature 518.7537 (2015):89–93.
13. Ma NJ, Isaacs FJ. Genomic recoding broadly obstructs the propagation of horizontally transferred genetic elements. Cell Syst. 2016;3:199–207.
14. Quax TEF, Claassens NJ, Söll D, van der Oost J. Codon bias as a means to fine-tune gene expression. Mol Cell. 2015;59:149–61.
15. Goodman DB, Kuznetsov G, Lajoie ML, Ahern BW, Napolitano MG, Chen KY et al. Millstone: software for multiplex microbial genome analysis and engineering. Genome Biol. 2017. doi:(10.1186/s13059-017-1223-1.
16. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6:80–92.
17. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006;2:2006.0008.
18. Zou H, Hui Z, Trevor H. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol. 2005;67:301–20.
19. Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF. Negative epistasis between beneficial mutations in an evolving bacterial population. Science. 2011;332:1193–6.
20. DiCarlo JE, Conley AJ, Penttilä M, Jäntti J, Wang HH, Church GM. Yeast oligo-mediated genome engineering (YOGE). ACS Synth Biol. 2013;2:741–9.
21. Kabadi AM, Ousterout DG, Hilton IB, Gersbach CA. Multiplex CRISPR/Cas9-based genome engineering from a single lentiviral vector. Nucleic Acids Res. 2014;42:e147.
22. Raman S, Rogers JK, Taylor ND, Church GM. Evolution-guided optimization of biosynthetic pathways. Proc Natl Acad Sci U S A. 2014;111:17803–8.
23. Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. Inexpensive multiplexed library preparation for megabase-sized genomes. PLoS One. 2015;10:e0128036.
24. Bonde MT, Kosuri S, Genee HJ, Sarup-Lytzen K, Church GM, Sommer MOA, et al. Direct mutagenesis of thousands of genomic targets using microarray-derived oligonucleotides. ACS Synth Biol. 2015;4:17–22.
25. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014;343:84–7.
26. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. Science. 2014;343:80–4.
27. Mansell TJ, Warner JR, Gill RT. Trackable multiplex recombineering for gene-trait mapping in E. coli. Methods Mol Biol. 2013;985:223–46.
28. Zeitoun RI, Garst AD, Degen GD, Pines G, Mansell TJ, Glebes TY, et al. Multiplexed tracking of combinatorial genomic mutations in engineered cell populations. Nat Biotechnol. 2015;33:631–7.
29. Garst AD, Bassalo MC, Pines G, Lynch SA, Halweg-Edwards AL, Liu R, Liang L, Wang Z, Zeitoun R, Alexander WG, Ryan T Gill. Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. Nat. Biotechnol. 2017;35:48–55.
30. Markham NR, Zuker M. UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol. 2008;453:3–31.
31. Salis HM. The ribosome binding site calculator. Methods Enzymol. 2011;498:19–42.
32. Gama-Castro S, Socorro G-C, Heladia S, Alberto S-Z, Daniela L-T, Luis M-R, et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res. 2015;44: D133–43.
33. Gregg CJ, Lajoie MJ, Napolitano MG, Mosberg JA, Goodman DB, Aach J, et al. Rational optimization of tolC as a powerful dual selectable marker for genome engineering. Nucleic Acids Res. 2014;42:4779–90.
34. Rohland N, Reich D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res. 2012;22:939–46.
35. Pedregosa F, et al. "Scikit-learn: Machine learning in Python." J. Mach. Learn. Res. 2011;2825-2830.
36. Lutz R, Bujard H. Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. Nucleic Acids Res. 1997;25:1203–10.
37. Sandberg TE, Pedersen M, LaCroix RA, Ebrahim A, Bonde M, Herrgard MJ, et al. Evolution of Escherichia coli to 42 C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. Mol Biol Evol. 2014;31:2647–62.
38. LaCroix RA, Sandberg TE, O'Brien EJ, Utrilla J, Ebrahim A, Guzman GI, et al. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. Appl Environ Microbiol. 2015;81:17–30.
39. Isaacs FJ, Carr PA, Wang HH, Lajoie MJ, Sterling B, Kraal L, et al. Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. Science. 2011;333:348–53.
40. Chou H-H, Chiu H-C, Delaney NF, Segrè D, Marx CJ. Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science. 2011; 332:1190–2.
41. Costantino N, Court DL. Enhanced levels of Red-mediated recombinants in mismatch repair mutants. Proc Natl Acad Sci. 2003;100:15748–53.
42. Donovan GT, Norton JP, Bower JM, Mulvey MA. Adenylate cyclase and the cyclic AMP receptor protein modulate stress resistance and virulence capacity of uropathogenic Escherichia coli. Infect Immun. 2013;81:249–58.

43. Raivio TL, Leblanc SKD, Price NL. The Escherichia coli Cpx envelope stress response regulates genes of diverse function that impact antibiotic resistance and membrane integrity. J Bacteriol. 2013;195:2755–67.
44. Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, Muñiz-Rascado L, et al. EcoCyc: a comprehensive database of Escherichia coli biology. Nucleic Acids Res. 2011;39:D583–90.
45. Ashburner M, Michael A, Ball CA, Blake JA, David B, Heather B, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000; 25:25–9.