ACS Synthetic Biology

Technical Note

pubs.acs.org/synthbio

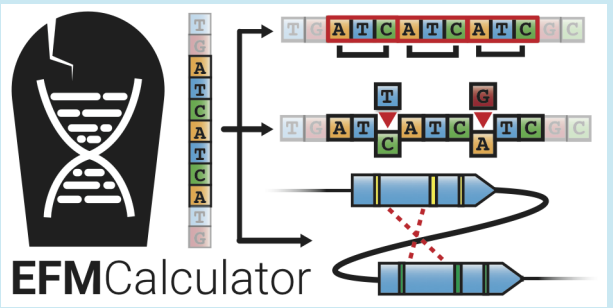
Predicting the Genetic Stability of Engineered DNA Sequences with the EFM Calculator

Benjamin R. Jack, Sean P. Leonard, Dennis M. Mishler, Brian A. Renda, Dacia Leon, Gabriel A. Suárez, and Jeffrey E. Barrick*

Center for Systems and Synthetic Biology, Center for Computational Biology and Bioinformatics, Institute for Cellular and Molecular Biology, Department of Molecular Biosciences, The University of Texas at Austin, Austin, Texas 78712, United States

Supporting Information

ABSTRACT: Unwanted evolution can rapidly degrade the performance of genetically engineered circuits and metabolic pathways installed in living organisms. We created the Evolutionary Failure Mode (EFM) Calculator to computationally detect common sources of genetic instability in an input DNA sequence. It predicts two types of mutational hotspots: deletions mediated by homologous recombination and indels caused by replication slippage on simple sequence repeats. We tested the performance of our algorithm on genetic circuits that were previously redesigned for greater evolutionary reliability and analyzed the stability of sequences in the iGEM Registry of Standard Biological Parts. More than half of the parts in the Registry are predicted to experience >100-fold elevated mutation rates due to the inclusion of unstable sequence configurations. We anticipate that the EFM Calculator will be a useful negative design tool for avoiding volatile DNA encodings, thereby increasing the evolutionary lifetimes of synthetic biology devices.



KEYWORDS: *computer-aided design (CAD), design-build-test cycle, genetic robustness, genetic engineering, hypermutable site, metabolic engineering*

Inactivating mutations can rapidly accumulate in the DNA sequences of synthetic biology devices, especially when an engineered activity imposes a fitness cost on the host cell.[1,2] These stochastic "evolutionary failure modes" (EFMs) decrease the predictability and productivity of bioengineering. Thus, designing robustness against evolutionary failure into a DNA sequence is an important goal for synthetic biology. Often, mutational hotspots in a DNA construct lead to a small number of very specific molecular events repeatedly arising and dominating the EFMs observed in malfunctioning copies of a device.[3,4]

Certain DNA sequence features are expected to be highly unstable in nearly any chassis organism. Homologous recombination between long repeat sequences commonly leads to gene-sized or larger deletions. The rates of these repeat-mediated deletions (RMDs) have been characterized extensively in bacterial plasmids.[5] Simple sequence repeats (SSRs) are hotspots for short insertion and deletion mutations. These consecutive copies of units consisting of one to a few DNA bases are inherently unstable due to polymerase slippage between units during DNA replication. Computational tools have been created for predicting the relative instabilities of different SSRs because these rapidly evolving sequences (also known as microsatellites and variable number tandem repeats) are often used in genotyping assays.[6]

Many different DNA sequences can potentially be used to specify equivalent biological devices—by altering codon usage in open-reading frames or by swapping out promoters of equal strength, for example. Although it would be beneficial to exclude sequence motifs that act as mutational hotspots when choosing how to encode a device in DNA, this step is not currently implemented in computer-aided design programs for synthetic biology. Here, we describe the first generation of the EFM Calculator, a software program that computationally predicts potential mutational vulnerabilities in an input DNA sequence so that they can be avoided.

RESULTS AND DISCUSSION

The EFM Calculator (http://barricklab.org/efm) is a web tool that predicts mutation rates at RMD and SSR sites and normalizes them to a baseline rate of point mutations to calculate a relative instability prediction (RIP) score for an input DNA sequence (see Methods) (Figure 1a). Sequences with RIP scores = 1 contain no predicted hypermutable sites. RIP scores >1 occur when SSR and RMD hotspots are detected in a sequence. By editing these high-frequency failure modes

Special Issue: IWBDA 2014

Received: April 7, 2015
Published: June 22, 2015

DOI: 10.1021/acssynbio.5b00068
ACS Synth. Biol. 2015, 4, 939−943

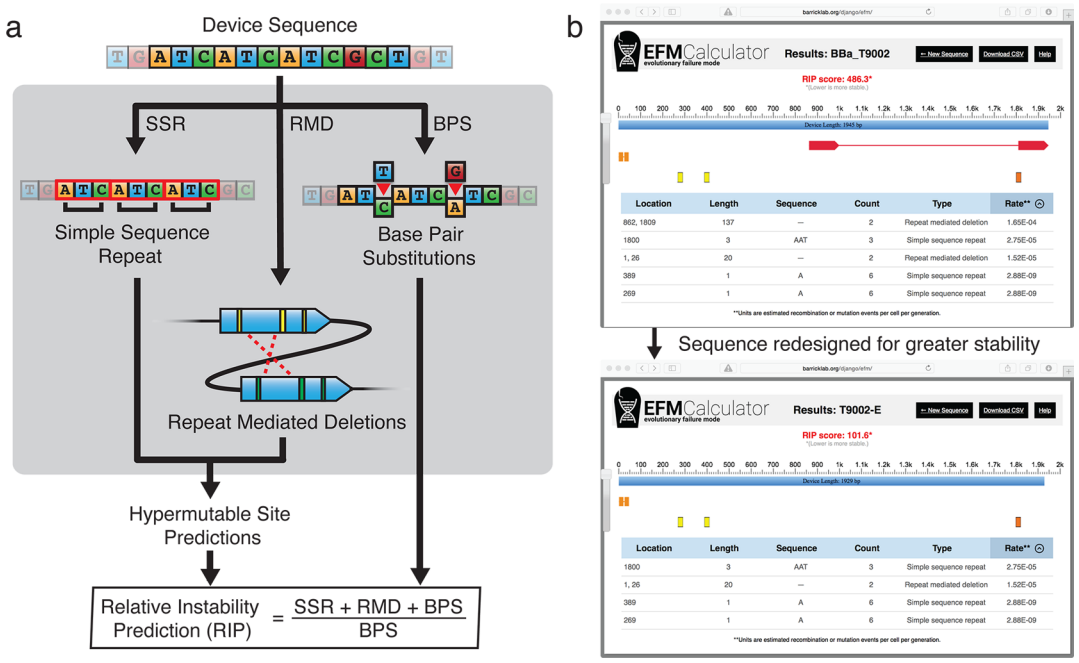


Figure 1. (a) The EFM Calculator accepts a DNA sequence as input, predicts two types of hypermutable sites in this sequence, and summarizes the overall prediction of instability with a RIP score. It outputs an HTML file of hypermutable sites on an interactive drawing of the sequence and in a table. The RIP score represents the factor by which redesigning the input sequence to eliminate predicted SSR and RMD mutational hotspots (leaving only the baseline rate of BPS mutations) could theoretically reduce the overall rate of mutations that contribute to the evolutionary failure modes of this DNA sequence when it is deployed in a bacterial host. (b) The original version of a genetic circuit that expresses GFP (BioBrick part T9002) receives a RIP score of 486.3. A redesigned version (T9002-E), shown experimentally to have a longer evolutionary half-life by Sleight et al.,[4] receives a more stable RIP score of 101.6.

out of a sequence, its mutation rate could theoretically be lowered by a factor equal to the RIP score. The EFM Calculator utilizes mutation rate models that are based on a generic wild-type bacterial host (e.g., *Escherichia coli*) by default, but it also allows the user to specify a *recA*$^-$ bacterial host (with lower rates of homologous recombination) for these calculations.

To test the EFM Calculator, we analyzed two sets of genetic circuits that had previously been redesigned for greater evolutionary robustness.[4] In this study, Sleight et al. measured an evolutionary half-life ($E_{1/2}$) for each plasmid-encoded circuit in *E. coli*, defined as the number of cell divisions (generations) before its GFP output in a cell population decayed to 50% of its original value.[3] Then, they profiled what mutations were causing the failure of each circuit by sequencing plasmids with malfunctioning devices isolated from independently evolved cells. We were careful to compare only redesigned circuits from Sleight et al. that maintained similar GFP expression levels, as this "cost" parameter was also found to greatly impact device half-lives.[4]

The first circuit (T9002) received a very high RIP score of 486.3 from the EFM Calculator. $E_{1/2}$ for this construct was only 7.1 generations (~1 day of laboratory culture).[4] Experimentally, T9002 always failed due to a deletion between two copies of the same terminator. The EFM Calculator was able to accurately predict that this RMD would be the dominant hotspot (Figure 1b). A redesigned version of this circuit (T9002-E) that altered one of the terminators has a lower RIP score of 101.6. In agreement with this prediction, T9002-E had an $E_{1/2}$ of ~16.7 generations and failed due to a variety of different mutations, including a different RMD involving operator sequences,[4] which was also predicted by the EFM Calculator.

A second circuit, I7101 (R0011 + E0240), initially had a RIP score of 180, dominated by a predicted RMD between *tetR* operator sequences. This construct had an $E_{1/2}$ of 19.8 generations, and all characterized mutants had this deletion.[4] A re-engineered version of I7101 (R0010 + E0240) with an alternative promoter had a reduced RIP of 120, and the experimentally measured $E_{1/2}$ doubled to ~42.4 generations. Interestingly, this circuit appeared to lose function due to unknown mutations elsewhere in the plasmid or host genome (i.e., outside of the circuit).[4]

To more widely examine how mutational hotspots might be affecting synthetic biology, we used the EFM Calculator to survey BioBrick sequences deposited in the International Genetically Engineered Machine (iGEM) Registry of Standard Biological Parts (http://parts.igem.org). Of the 20,952 parts with lengths of >50 base pairs in the iGEM Registry, 57% had RIP scores >100, and 8.5% had RIP scores >1000 when using the default *recA*$^+$ bacterial host setting (Figure 2a). Control data sets in which the DNA sequences for each part in the Registry were randomly shuffled (see Methods) exhibited much lower levels of predicted instability. On average, only 19% of the parts in each shuffled data set had RIP scores >100, and only 0.38% had RIP scores >1000. In fact, none of the control data sets had as many parts with RIP scores above these cutoffs as were in the actual Registry (one-tailed Monte Carlo test, $p < 0.001$). This analysis suggests that the genetic stability of many parts in the iGEM Registry could be substantially improved by redesigning their sequences to eliminate mutational hotspots predicted by the EFM Calculator.

DOI: 10.1021/acssynbio.5b00068
ACS Synth. Biol. 2015, 4, 939−943

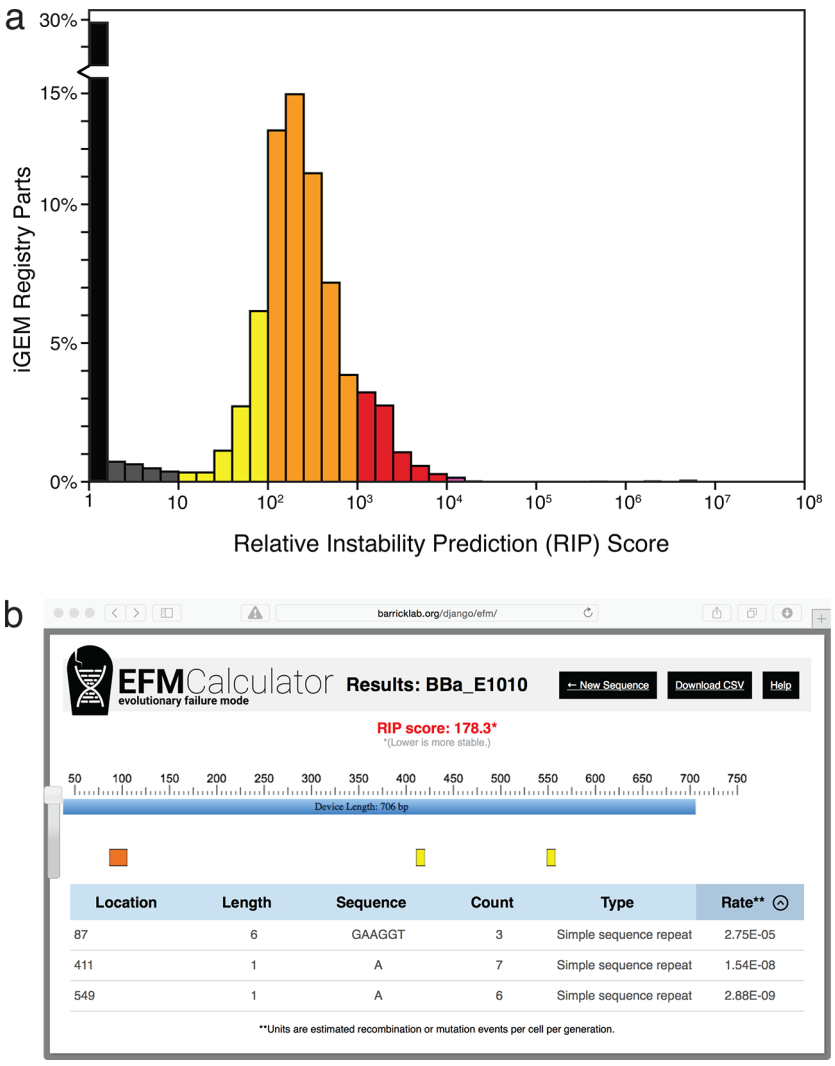


Figure 2. (a) Distribution of RIP scores predicted by the EFM Calculator for parts with lengths of >50 base pairs in the iGEM Registry. The colored shading of bars demarcates 10-fold increases in predicted mutational instability. (b) EFM Calculator output for BioBrick part E1010, one of the top ten most used coding sequences in the Registry. The RIP score of 178.3 indicates that the mutation rate in this part is predicted to be 178 times the baseline rate of base-pair substitutions. Three simple sequence repeats contribute to this genetic instability prediction.

Of the top ten most frequently used coding region parts, three were predicted to have greatly elevated mutation rates: C0061, a *luxI* autoinducer synthetase (RIP 226); C0012, a *lacI* repressor with a degradation tag (RIP 218); and E1010, an engineered mutant of red fluorescent protein (RIP 178.3). All three parts were predicted to be unstable due to SSRs. For E1010, these included a hexanucleotide triplet repeat and two mononucleotide runs (Figure 2b). As was the case for these sequences, the EFM Calculator predicted that SSRs also dominate the mutation rates for 70.8% of all parts in the Registry with RIP scores >100. The other 29.2%, with mutation rates dominated by RMDs, were often composite parts (i.e., devices consisting of multiple subparts) that contained homologous repeats due to reusing exact copies of the same genetic part, like the original versions of the circuits that were studied by Sleight et al.[4]

The extent to which the mutational hotspots predicted by the EFM Calculator impact the overall reliability of synthetic biology across different organisms and applications is unknown. Researchers rarely characterize why specific DNA sequences fail to function as planned; a trial-and-error approach of creating several constructs and only studying whichever one functions best is more typical. Thus, although the quantitative predictions of the EFM Calculator are based on experimentally determined mutation rates, there are very few measurements of the evolutionary half-lives of genetic devices with which to test whether these mutations lead to important EFMs. We hope that making this tool available online will encourage more synthetic biologists and iGEM teams to consider and characterize the genetic stability of their DNA designs.

The SSR and RMD mutations predicted by the EFM Calculator occur with greatly elevated rates in most organisms due to the highly conserved mechanisms of DNA replication and repair.[2] Still, the overall evolutionary reliability of any given DNA-encoded device may also depend critically on other factors. For example, transposable genetic elements represent a competing mutational process; they can inactivate synthetic constructs when new transposon copies insert into critical genes.[7] However, rates of transposition vary greatly, even among closely related *E. coli* strains, and many transposon families do not have target sites with any predictable DNA sequence conservation.[8] Thus, it would be difficult to

DOI: 10.1021/acssynbio.5b00068
ACS Synth. Biol. 2015, 4, 939−943

incorporate predictions of transposon mutations into the EFM Calculator. Instead, the best way to address this source of genetic instability may be to use "clean-genome" microbial strains in which this mutation type has been eliminated by deleting all transposable elements from the host genome.[7]

Whether an inactivating mutation in an engineered DNA device affects the function of the cell population on a relevant time scale depends not only on the rate at which it arises but also on the fitness benefit of encoding an inactivated device variant relative to an unmutated device.[2] Part of the cost of engineered DNA devices is simply due to expressing additional protein components. This metabolic burden is influenced by many factors (promoter strength, ribosome binding site, codon usage, plasmid copy number), and a higher burden is associated with decreased overall performance and stability.[9] Some of these gene expression properties can be predicted by tools, such as the Ribosome Binding Site Calculator,[10] but there are also fitness costs that are specific to how different heterologous proteins and pathways may interfere with host cell replication. The complex, emergent nature of these interactions makes them very difficult to accurately predict from DNA sequence alone, so they are also outside the scope of the EFM calculator.

There are additional hypermutable DNA sequence features that could be incorporated into future versions of the EFM calculator to further refine RIP score calculations. For example, intrastrand DNA secondary structures can expose specific bases to elevated rates of chemical damage.[2] However, even at this stage, the EFM Calculator can be used as an effective negative design tool. By alerting users to the presence of putative hypermutable sites in their DNA constructs, they can at least avoid these known sources of genetic instability that are intrinsic to the sequence of the part alone. The EFM Calculator provides scientists and engineers with the means to redesign potentially volatile sequences *in silico* before they waste time and effort creating genetic devices that are unlikely to function reliably when deployed in living organisms.

METHODS

EFM Calculator Implementation. The EFM Calculator (version 1.0.0) is a web tool implemented in Python with the Django framework. DNA sequences are input in FASTA, GenBank, or XML format using Biopython.[11] Processing occurs via Python code that invokes MUMmer (version 3.23).[12] An interactive HTML file with JavaScript generated using the SCRIBL library[13] and a machine-readable comma-separated values (CSV) text file are output. For input sequences in GenBank or XML formats, there is an option to include only mutations predicted within regions of the sequence that are annotated with features (e.g., protein-coding genes and promoters) in the RIP calculations and output.

Simple Sequence Repeat (SSR) Predictions. SSRs are defined by two parameters: the length of the repeat unit (L), in base pairs, and the number of consecutive copies of the repeat unit (N) (Figure S1a, Supporting Information). The EFM Calculator identifies all potentially hypermutable SSRs with $L \leq 15$. Larger repeat units are more likely to lead to recombination than to polymerase slippage,[5] so they are accounted for by the RMD model described in the next section. SSRs must have a minimum length and repeat number to act as mutational hotspots.[14−17] Therefore, we only include mutation rate predictions for SSRs in which $N \geq 4$ for the case of $L = 1$ and $N \geq 3$ for the case of $L \geq 2$.

Previous analyses have established an exponential relationship between values of N and the per-generation mutation rate (μ),[6] so we used this as a basis for our quantitative modeling. For each SSR, the EFM Calculator estimates a per-generation mutation rate (μ) according to one of the following two models: $\log_{10}(\mu) = -12.9 + 0.729N$ ($L = 1$) and $\log_{10}(\mu) = -4.749 + 0.063N$ ($L \geq 2$). We determined these values from log−linear fits to published experimental data for the cases of $L = 1$ ($R^2 = 0.95$, $p = 0.003$) and $L \geq 2$ ($R^2 = 0.37$, $p = 4 \times 10^{-8}$) (Figure S1b, Supporting Information). For $L = 1$, we fit data from a mutation accumulation experiment with *E. coli*.[14] For $L \geq 2$, we fit data from targeted sequencing of specific loci during parallel serial passage experiments performed on *E. coli*,[15] *Burkholderia pseudomallei*,[16] and *Yersinia pestis*.[17] When multiple SSRs overlap, the EFM Calculator includes only the dominant one (with maximum N and minimum L).

Repeat-Mediated Deletion (RMD) Predictions. Hypermutable direct sequence repeats are defined by two parameters: length of each copy of the repeat (L_R), in base pairs, and the distance between the repeats (D), in base pairs (Figure S2a, Supporting Information). The EFM Calculator uses MUMmer to identify all exact repeats with $L_R \geq 16$ in the input sequence. Only repeats with the same orientation (for which a recombination event will lead to a deletion) are considered in mutation rate calculations. Shorter repeats and repeat sequences interrupted by base mismatches generally experience greatly reduced recombination rates,[5] so we do not include these cases in our model. The L and D values for each direct sequence repeat are used to predict an RMD rate according to empirical equations (specific to *recA*$^-$ or *recA*$^+$ hosts) constructed by Oliveira et al. from a meta-analysis of recombination rates observed on multicopy plasmids in *E. coli* and *B. subtilis*[5] (Figure S2b,c, Supporting Information). These models have also been reported to predict genomic recombination rates in other bacteria reasonably well.[5] If there are more than two copies of the same repeat, then the rates for every pair are added together to estimate the total predicted RMD rate.

Relative Instability Prediction (RIP) Score. The EFM Calculator outputs a consolidated RIP score that reflects the overall instability of an input DNA sequence relative to a sequence of the same length that does not include any predicted mutational hotspots. For this baseline, we used the rate of base-pair substitution (BPS) mutations estimated from genome sequencing of *E. coli* mutation accumulation lines.[14] The expected baseline BPS mutation rate in an input DNA sequence is equal to this per-base rate times the length of the sequence. The RIP score is defined as the sum of the predicted RMD, SSR, and BPS rates for a sequence divided by the BPS rate.

Sequences with RIP scores = 1 have no predicted SSR or RMD mutations. Sequences with RIP scores >1 include SSR and/or RMD sites that are expected to lead to increased rates of device failure. For example, a RIP score of 100 means that the sequence is predicted to experience a total rate of SSR and RMD mutations that is 99 times the baseline BPS mutation rate alone. Therefore, the RIP score is also an estimate of the greatest factor by which one can expect to improve the genetic stability of a given device by redesigning its sequence to eliminate all predicted mutational hotspots.

iGEM Registry Analysis. Sequences were retrieved from the iGEM Registry Web site as a single FASTA file that included all parts submitted before July 29, 2013. Only parts

DOI: 10.1021/acssynbio.5b00068
ACS Synth. Biol. 2015, 4, 939−943

longer than 50 base pairs in length were analyzed. We generated 956 control data sets with the exact same base compositions, lengths, and sequence counts as the Registry by randomly shuffling the sequences of every part. We compared the distributions of RIP scores predicted by the EFM Calculator for these Monte Carlo randomized data sets to the results for the iGEM Registry sequences to determine whether the actual parts were significantly less stable than expected.

■ ASSOCIATED CONTENT

Supporting Information

Figures showing the SSR and RMD mutation rate models used by the EFM Calculator. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acssynbio.5b00068.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jbarrick@cm.utexas.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Barrick lab colleagues and two anonymous reviewers for helpful suggestions. This research was supported by the U.S. National Institutes of Health (R00-GM087550), U.S. National Science Foundation BEACON Center for the Study of Evolution in Action (DBI-0939454), U.S. Army Research Office (W911NF-12-1-0390), and the Cancer Prevention & Research Institute of Texas (RP130124).

■ REFERENCES

(1) Arkin, A. P., and Fletcher, D. A. (2006) Fast, cheap and somewhat in control. *Genome Biol. 7*, 114.

(2) Renda, B. A., Hammerling, M. J., and Barrick, J. E. (2014) Engineering reduced evolutionary potential for synthetic biology. *Mol. BioSyst. 10*, 1668−1678.

(3) Canton, B., Labno, A., and Endy, D. (2008) Refinement and standardization of synthetic biological parts and devices. *Nat. Biotechnol. 26*, 787−793.

(4) Sleight, S. C., Bartley, B. A., Lieviant, J. A., and Sauro, H. M. (2010) Designing and engineering evolutionary robust genetic circuits. *J. Biol. Eng. 4*, 12.

(5) Oliveira, P. H., Lemos, F., Monteiro, G. A., and Prazeres, D. M. F. (2008) Recombination frequency in plasmid DNA containing direct repeats—predictive correlation with repeat and intervening sequence length. *Plasmid 60*, 159−165.

(6) Legendre, M., Pochet, N., Pak, T., and Verstrepen, K. J. (2007) Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res. 17*, 1787−1796.

(7) Pósfai, G., Plunkett, G., Fehér, T., Frisch, D., Keil, G. M., Umenhoffer, K., Kolisnychenko, V., Stahl, B., Sharma, S. S., de Arruda, M., Burland, V., Harcum, S. W., and Blattner, F. R. (2006) Emergent properties of reduced-genome *Escherichia coli*. *Science 312*, 1044−1046.

(8) Sousa, A., Bourgard, C., Wahl, L. M., and Gordo, I. (2013) Rates of transposition in *Escherichia coli*. *Biol. Lett. (London, U.K.) 9*, 20130838.

(9) Ceroni, F., Algar, R., Stan, G.-B., and Ellis, T. (2015) Quantifying cellular capacity identifies gene expression designs with reduced burden. *Nat. Methods 12*, 415−418.

(10) Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol. 27*, 946−950.

(11) Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and De Hoon, M. J. L. (2009) Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics 25*, 1422−1423.

(12) Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004) Versatile and open software for comparing large genomes. *Genome Biol. 5*, R12.

(13) Miller, C. A., Anthony, J., Meyer, M. M., and Marth, G. (2013) Scribl: an HTML5 Canvas-based graphics library for visualizing genomic data over the web. *Bioinformatics 29*, 381−383.

(14) Lee, H., Popodi, E., Tang, H., and Foster, P. L. (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci. U.S.A. 109*, E2774−E2783.

(15) Vogler, A. J., Keys, C., Nemoto, Y., Colman, R. E., Jay, Z., and Keim, P. (2006) Effect of repeat copy number on variable-number tandem repeat mutations in *Escherichia coli* O157:H7. *J. Bacteriol. 188*, 4253−4263.

(16) U'Ren, J. M., Schupp, J. M., Pearson, T., Hornstra, H., Friedman, C. L. C., Smith, K. L., Daugherty, R. R. L., Rhoton, S. D., Leadem, B., Georgia, S., Cardon, M., Huynh, L. Y., DeShazer, D., Harvey, S. P., Robison, R., Gal, D., Mayo, M. J., Wagner, D., Currie, B. J., and Keim, P. (2007) Tandem repeat regions within the *Burkholderia pseudomallei* genome and their application for high resolution genotyping. *BMC Microbiol. 7*, 23.

(17) Girard, J. M., Wagner, D. M., Vogler, A. J., Keys, C., Allender, C. J., Drickamer, L. C., and Keim, P. (2004) Differential plague-transmission dynamics determine *Yersinia pestis* population genetic structure on local, regional, and global scales. *Proc. Natl. Acad. Sci. U.S.A. 101*, 8408−8413.

DOI: 10.1021/acssynbio.5b00068
ACS Synth. Biol. 2015, 4, 939−943