SYNTHETIC BIOLOGY

Synthetic Biology, 2022, **7(1)**, 1–6
DOI: https://doi.org/10.1093/synbio/ysac006
Advance access publication date: 28 April 2022
Software

Highly efficient libraries design for saturation mutagenesis

Gur Pines[1,*], Assaf Pines, and Carrie A. Eckert[2,*]

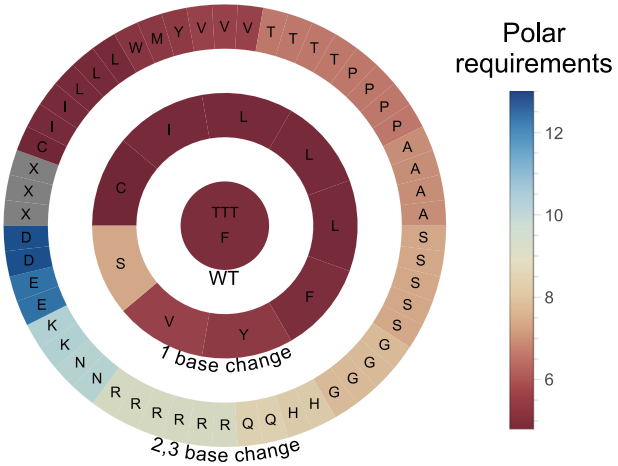
[1]Department of Entomology, Institute of Plant Protection, Agricultural Research Organization—The Volcani Center, Rishon, LeZion, Israel
[2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
*Corresponding author: E-mail: gurp@volcani.agri.gov.il eckertca@ornl.gov

Abstract

Saturation mutagenesis is a semi-rational approach for protein engineering where sites are saturated either entirely or partially to include amino acids of interest. We previously reported on a codon compression algorithm, where a set of minimal degenerate codons are selected according to user-defined parameters such as the target organism, type of saturation and usage levels. Here, we communicate an addition to our web tool that considers the distance between the wild-type codon and the library, depending on its purpose. These forms of restricted collections further reduce library size, lowering downstream screening efforts or, in turn, allowing more comprehensive saturation of multiple sites. The library design tool can be accessed via http://www.dynamcc.com/dynamcc_d/.

Key words: saturation mutagenesis; protein engineering; library design; codon compression

Graphical Abstract



1. Introduction

Saturation mutagenesis is used when the target amino acids for mutagenesis are known, but their final identity is yet to be determined. Hence, these target sites are mutated to include all possible amino acids, and the mutants are screened or selected for the desired phenotype (1). When more knowledge on the target sites is available, restricted saturation is employed, limiting the saturation to a subset of selected amino acids. When saturating several sites in combination, the library size increases exponentially, restricting the number of explorable combinations, especially when using screening methods (rather than selection) for the desired phenotype (2–4). For example, when saturating three

Submitted: 18 April 2021; **Received (in revised form):** 24 April 2022; **Accepted:** 11 May 2022

sites using the common NNK codon (N = any nucleotide, K = G or T. The NNK codon covers 32 codons that include all amino acids albeit with redundancy and a stop codon), the number of variants needed to be screened to achieve 95% library coverage reaches 98 164. However, when using 20 codons, allowing complete saturation (including wild-type amino acids for combinatorial purposes), the number is reduced to 23 966 (Equation 1).

$$T = -v^s \cdot ln(1-p) \tag{1}$$

where T represents the library size or colonies needed to be screened, V is the number of codon variants (32 and 20 in the example above, respectively), s is the number of combinations (equal to 3 here) and p represents the desired library coverage, which is set to 0.95 in our calculation (2, 5, 6).

Hence, it is preferable to reduce the library size wherever possible. Due to the random nature of naturally occurring mutations, it is extremely rare to achieve more than a single-base change within a codon (7). Therefore, studying naturally occurring mutations such as recapitulating evolutionary processes or estimating mutants' oncogenicity may benefit from addressing single-base changes only. These single-base differences are defined as having a hamming distance of 1. Such an approach will reduce the number of codons from 32 (when using the common NNK codon) to just 9. In contrast, it may be favorable for protein engineering purposes to focus on hamming distance that is larger than 1, i.e. having two or three base changes. It has been shown that in many cases, the best-performing mutants require more than a single-base change as such codon distance allows exploring a broader diversity of

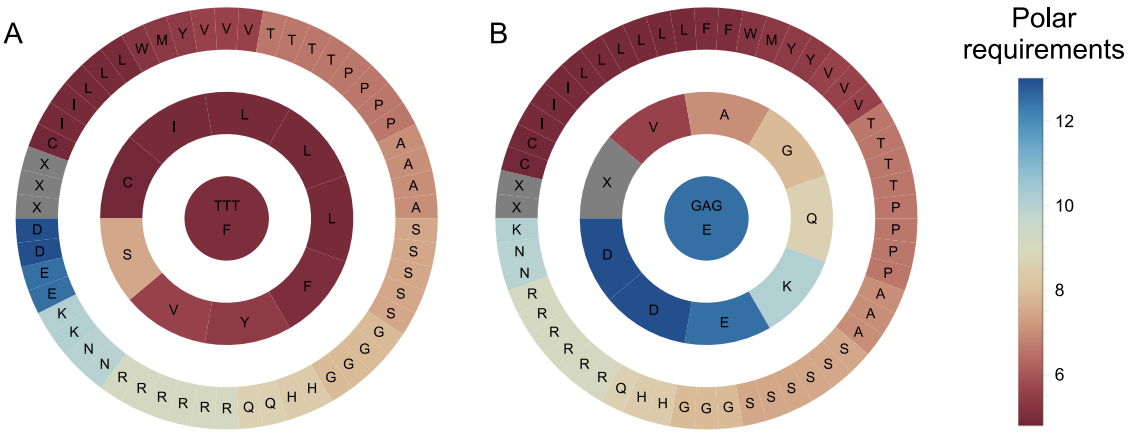


Figure 1. Codon accessibility examples. The center circle denotes the wild-type codon. The inner ring displays the accessible amino acids via a single-base substitution, and the outer ring represents the two- and three-base change accessibility. Colors correlate with amino acid polar requirements. A. A single-base change within the TTT codon does restrict accessibility to amino acids with low polar requirements and thus does not explore the complete possible amino acid repertoire. Removing the inner ring codons and focusing on codon distance larger than 1 results in a much more balanced library with amino acids spanning across the complete polar requirements spectrum. B. The GAG codon represents a more evolvable codon, representing a variety of amino acids in both rings.

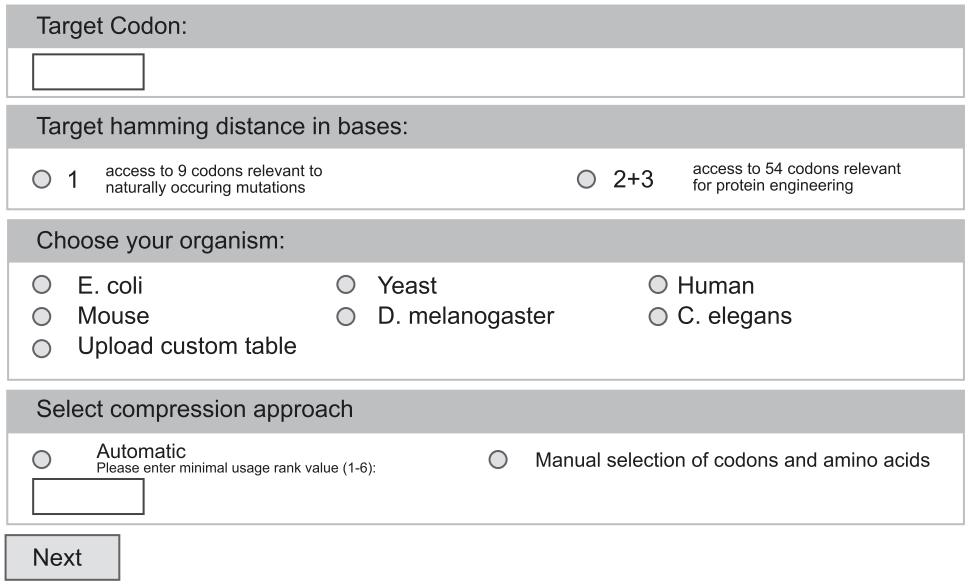


Figure 2. The DYNAMCC_D screen. Users are requested to input the wild-type codon, the base distance the library should have from the wild-type codon, the target organism and the compression approach. Depending on the compression approach, hitting the Next button will either lead to the result page (automatic option) or to the codon and amino acid selection (Manual selection, Figure 3).

amino acids, a property defined by the structure of the genetic code (8–22). The genetic code is robust to mutation while still allowing a degree of evolvability, and when restricting a codon mutation to a single base, there is, on average, a 40% chance that the mutated codon will code for an identical or chemically similar amino acid (8, 12, 23). For example, when examining amino acid polar requirements as a metric (24), it is clear that while individual codons possess different evolvability potential (Figure 1A and B), most codons will have access to chemically diverse amino acids when removing the single-nucleotide polymorphism (SNP) accessible codons (Fig. S1). For example, unlike the GAG codon, SNPs in the TTT cannot explore amino acids with distant polar requirements values. However, both codons span a wide polar requirement coverage when looking at the non-SNP codons (Figure 1, outer rings). A complete accessibility map of all 64 codons is available as Supplementary Figure S1.

2. DYNAMCC_D

2.1. The Dynamic Management of Codon Compression tools for saturation mutagenesis

We previously reported on the Dynamic Management of Codon Compression (DYNAMCC) approach that enables the user to remove undesired elements from the saturation library and outputs a minimal list of compressed codons using the IUPAC nucleic acid notation (25, 26). We denote this action as 'codon compression' while the opposite action of deriving individual codons from a compressed codon is termed as 'codon explosion'. Undesired library elements may include stop codons, the wild-type amino acid and redundancy, which exist in the commonly used NNK codon. The DYNAMCC_0 tool allows the user to remove redundancy, where amino acids are coded for by multiple codons within the library. In addition, it is possible to avoid stop codons and the wild-type amino acid. Moreover, the algorithm takes into account

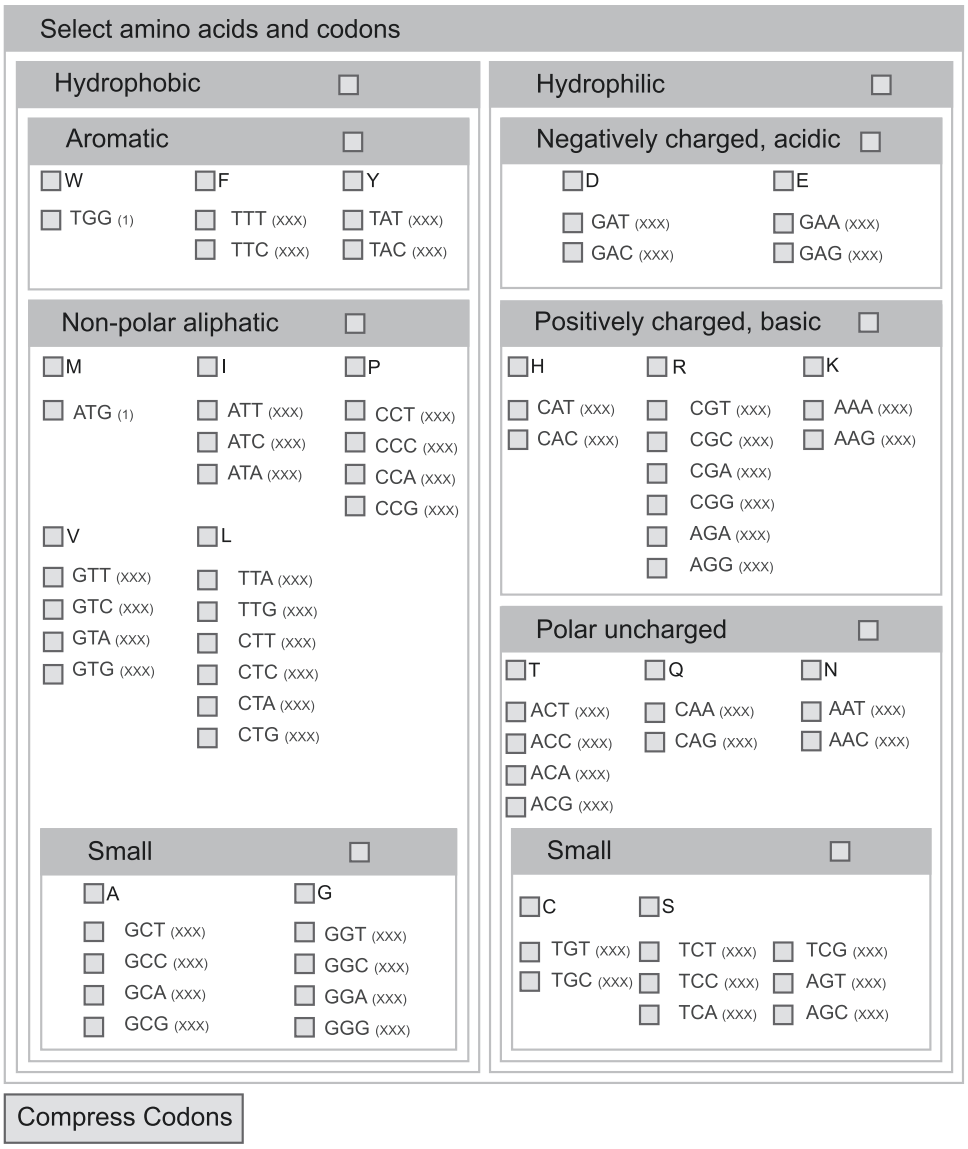


Figure 3. Codon and amino acid selection screen. When the Manual Selection is selected on the first screen (Figure 2), the available amino acids and codons are clickable, whereas the nonavailable codons are nonclickable and gray shaded. Amino acids are ordered according to their properties to allow easy identification and batch selection. By default, the codons with the highest usage value are preselected for each possible amino acid. XXX denotes codon usage, and values depend on the organism of choice.

the codon usage values for the destination organism to ensure that the library will include suitable codons. In the case of a codon usage table not preloaded to the website, the user can upload the table of their organism of interest to allow maximal flexibility. The DYNAMCC_R tool is designed for experiments where the redundant space is of interest and includes every codon for a selected amino acid in the library. This tool may be of interest when synonymous mutations may affect several factors such as messenger RNA stability, protein expression, folding and function (27–30). Both tools support restricted saturation as well, for instances where only a subset of amino acids are of interest. While the DYNAMCC tools provide better control over the library size, the tradeoff is that the number of compressed codons is larger than 1 as in the case of NNK. Hence, it balances using a single codon, with its redundancies on one end and synthesizing the complete codon collection on the other end.

2.2. Workflow and examples

Here, we expand our tool to take into account the distance between the wild-type and the library codons. We mark this new tool with the letter D for distance. The rationale for this tool's development is that libraries focusing on a single-base change or on larger distances tend to answer different questions, such as studying random mutagenesis versus protein engineering. Similar to other DYNAMCC tools, this tool results in achieving a compressed codon list that includes the desired codons.

Since base distance is defined at the codon level, the user is first required to input the wild-type codon as opposed to the other DYNAMCC tool that requires the wild-type amino acid (Figure 2). Then, the user should define the type of the desired library. As discussed above, single-base polymorphism is dominant when studying naturally occurring mutations. These libraries will result in access to 9 codons. The number of unique amino acids is

A Targeted codon: ATG (M)
Target hamming distance: 1
Selected organism usage table: E. coli

Compressed codon	Exploded codons	Rank	Usage	Amino acid
ATT	ATT	1	0.49	I
STG	CTG	1	0.47	L
	GTG	1	0.35	V
AVG	AAG	1	0.26	K
	ACG	1	0.25	T
	AGG	1	0.04	R

B Targeted codon: ATG (M)
Target hamming distance: 2+3
Selected organism usage table: E. coli

Compressed codon	Exploded codons	Rank	Usage	Amino acid
BDT	CAT	1	0.57	H
	CGT	1	0.36	R
	CTT	2	0.12	L
	GAT	1	0.63	D
	GGT	2	0.35	G
	GTT	1	0.28	V
	TAT	1	0.59	Y
	TGT	2	0.46	C
	TTT	1	0.58	F
SCG	CCG	1	0.49	P
	GCG	1	0.33	A
VAA	AAA	1	0.74	K
	CAA	2	0.34	Q
	GAA	1	0.68	E
TGG	TGG	1	1	W
AVC	AAC	1	0.51	N
	ACC	1	0.4	T
	AGC	1	0.25	S

Figure 4. DYNAMCC_D output. Results shown are for the saturation of methionine. A. Results with a distance of a single nucleotide from the wild-type codon. Such libraries are ideal for the study of naturally occurring mutations. B. Results of a query for two and three base differences. This type of library may be beneficial when a significant contrast in chemical properties is needed. In both examples, the automatic compression option was selected.

codon-dependent and can range between 5 and 8, with the rest of the codons counting for synonymous codons and stop codons. When increased accessibility to amino acids with different chemical properties is needed, as in some protein engineering efforts, two or more base changes may be required (Figure 1). These libraries result in accessibility to 54 codons, which are biased toward more substantial differences from the wild-type amino acid.

The third step requires the definition of the target organism, as in the other DYNAMCC tools, to ensure the selection of highly used codons. Similar to the other DYNAMCC tools, users may upload usage tables of non-model organisms. Finally, a compression approach should be selected. The automatic approach is similar to the process used in DYNAMCC_0, where the user defines the usage rank threshold for compression. These values range from 1 to 6, with the value of 1 restricting the algorithm to consider only the most highly used codons for each amino acid. The higher the value, the larger the codon pool to achieve efficient compression, which also correlates with increased computation time. We recommend not exceeding the value of 3, as this might increase the computation time to a point where the server returns an error. For example, when selecting an SNP library for the ATG codon, the resulting compressed codons are STG (codes for L and V) and AVG (K, T and R). To complete the library an additional uncompressed codon, ATT (I) is added (Figure 4A).

The second compression strategy provides the user with complete control over the selected codons for compression. Selecting this approach will direct the user to a second screen where all possible codons may be selected, with preselected highly used codons (Figure 3). This setting allows the removal of unwanted amino acids or focusing on several amino acids, including all possible redundancies. It should be noted that since the algorithm cannot select codons for efficient compression as in the previous strategy, this approach might result in libraries composed of larger pools.

The final library design is displayed on the results screen in a table format, similar to the other DYNAMCC tools (Figure 4).

3. Discussion

DYNAMCC_D is designed to increase the efficiency of saturation mutagenesis experiments. Our previous efforts focused on controlling the library size by removing undesired codons. Smaller libraries result in lower downstream efforts or allow to increase the search space by looking at additional sites. Here, we describe DYNAMCC_D, which takes into account the hamming distance between the wild-type codon and the library codons. SNPs are prevalent for randomly generated mutations, both naturally occurring and in laboratory settings. Hence, mutagenesis studies may benefit from significantly reducing library size since only nine codons are accessible via SNP, some of which may be synonymous or code for stop codons (Figure 1). Higher hamming distance correlates with a more extreme difference in chemical properties, which may be required in many protein engineering cases. Thus, limiting a library to a larger distance than 1 may not only reduce its size but may also enhance the contrast between beneficial and nonbeneficial mutations since such libraries are mainly enriched in deleterious mutations (12).

Supplementary data

Supplementary data are available at SYNBIO online.

Data availability

This tool is available at http://www.dynamcc.com/dynamcc_d/.

The DYANMCC algorithms were written in Python 2.7. Currently, the www.dynamcc.com domain is hosted at www.pythonanywhere.com.

The DYNAMCC_D website was written to provide an accessible library design tool to biologists with no computational background. However, similar to our previous DYNAMCC algorithms, this code is also freely available under the BSD 3-clause license, allowing modifications and code redistribution. The source code can be found at https://github.com/bioverse/dynamcc.

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-SC0018368.

Acknowledgments

We thank QF Network for its website programming help and Andrea Edwards for assistance with the website domain.

Conflict of interest statement. None declared.

References

1. Schultz,S.C. and Richards,J.H. (1986) Site-saturation studies of beta-lactamase: production and characterization of mutant beta-lactamases with all possible amino acid substitutions at residue 71. *Proc. Natl. Acad. Sci. U. S. A.*, **83**, 1588–1592.
2. Patrick,W.M. and Firth,A.E. (2005) Strategies and computational tools for improving randomized protein libraries. *Biomol. Eng.*, **22**, 105–112.
3. Reetz,M.T., Kahakeaw,D. and Lohmer,R. (2008) Addressing the numbers problem in directed evolution. *Chembiochem*, **9**, 1797–1804.
4. Pines,G. and Gill,R.T. (2018) Dynamic management of codon compression for saturation mutagenesis. In: Braman JC (ed). *Synthetic Biology: Methods and Protocols*. Springer, New York, pp. 171–189.
5. Reetz,M.T. (2011) Laboratory evolution of stereoselective enzymes: a prolific source of catalysts for asymmetric reactions. *Angew. Chem. Int. Ed Engl.*, **50**, 138–174.
6. Bosley,A.D. and Ostermeier,M. (2005) Mathematical expressions useful in the construction, description and evaluation of protein libraries. *Biomol. Eng.*, **22**, 57–61.
7. Smith,N.G.C., Webster,M.T. and Ellegren,H. (2003) A low rate of simultaneous double-nucleotide mutations in primates. *Mol. Biol. Evol.*, **20**, 47–53.
8. Miyazaki,K. and Arnold,F.H. (1999) Exploring nonnatural evolutionary pathways by saturation mutagenesis: rapid improvement of protein function. *J. Mol. Evol.*, **49**, 716–720.
9. Pines,G., Oh,E.J., Bassalo,M.C., Choudhury,A., Garst,A.D., Fankhauser,R.G., Eckert,C.A., Gill,R.T. *et al.* (2018) Genomic Deoxyxylulose Phosphate Reductoisomerase (DXR) mutations conferring resistance to the antimalarial drug fosmidomycin in *E. coli*. *ACS Synth. Biol*, **7**, 2824–2832.
10. Garst,A.D., Bassalo,M.C., Pines,G., Lynch,S.A., Halweg-Edwards,A.L., Liu,R., Liang,L., Wang,Z., Zeitoun,R., Alexander,W.G. *et al.* (2017) Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. *Nat. Biotechnol.*, **35**, 48–55.

11. Hu,S., Huang,J., Mei,L., Yu,Q., Yao,S., Jin,Z. *et al.* (2010) Altering the regioselectivity of cytochrome P450 BM-3 by saturation mutagenesis for the biosynthesis of indirubin. *J. Mol. Catal. B Enzym.*, **67**, 29–35.
12. Firnberg,E. and Ostermeier,M. (2013) The genetic code constrains yet facilitates Darwinian evolution. *Nucleic Acids Res.*, **41**, 7420–7428.
13. Rui,L., Reardon,K.F. and Wood,T.K. (2005) Protein engineering of toluene ortho-monooxygenase of Burkholderia cepacia G4 for regiospecific hydroxylation of indole to form various indigoid compounds. *Appl. Microbiol. Biotechnol.*, **66**, 422–429.
14. Fan,A. and Li,S.-M. (2016) Saturation mutagenesis on Arg244 of the tryptophan C4-prenyltransferase FgaPT2 leads to enhanced catalytic ability and different preferences for tryptophan-containing cyclic dipeptides. *Appl. Microbiol. Biotechnol.*, **100**, 5389–5399.
15. Xu,X., Chen,J., Wang,Q., Duan,C., Li,Y., Wang,R., Yang,S. *et al.* (2016) Mutagenesis of key residues in the binding center of l- aspartate-β-semialdehyde dehydrogenase from *Escherichia coli* enhances utilization of the cofactor NAD(H). *Chembiochem*, **17**, 56–64.
16. Qiu,Y.-L., He,Q.-H., Xu,Y., Wang,W. and Liu,-Y.-Y. (2016) Modification of a deoxynivalenol-antigen-mimicking nanobody to improve immunoassay sensitivity by site-saturation mutagenesis. *Anal. Bioanal. Chem.*, **408**, 895–903.
17. Cahn,J.K.B., Baumschlager,A., Brinkmann-Chen,S. and Arnold,F.H. (2016) Mutations in adenine-binding pockets enhance catalytic properties of NAD(P)H-dependent enzymes. *Protein Eng. Des. Sel.*, **29**, 31–38.
18. Asad,S., Dastgheib,S.M.M. and Khajeh,K. (2016) Construction of a horseradish peroxidase resistant toward hydrogen peroxide by saturation mutagenesis. *Biotechnol. Appl. Biochem.*, **63**, 789–794.
19. White,P.J., Squirrell,D.J., Arnaud,P., Lowe,C.R. and Murray,J.A. (1996) Improved thermostability of the North American firefly luciferase: saturation mutagenesis at position 354. *Biochem. J.*, **319**, 343–350.
20. Cormack,B.P., Valdivia,R.H. and Falkow,S. (1996) FACS-optimized mutants of the green fluorescent protein (GFP). *Gene*, **173**, 33–38.
21. Firnberg,E., Labonte,J.W., Gray,J.J., Ostermeier,M. and Comprehensive,A. (2014) High-resolution map of a Gene's fitness landscape. *Mol. Biol. Evol.*, **31**, 1581–1592.
22. Whitehead,T.A., Chevalier,A., Song,Y., Dreyfus,C., Fleishman,S.J., De Mattos,C., Myers,C.A., Kamisetty,H., Blair,P., Wilson,I.A. *et al.* (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.*, **30**, 543–548.
23. Pines,G., Winkler,J.D., Pines,A. and Gill,R.T. (2017) Refactoring the genetic code for increased evolvability. *MBio*, **8**, e01654–17.
24. Alff-Steinberger,C. (1969) The genetic code and error transmission. In: *Proc. Natl. Acad. Sci. U. S. A.*, **64**, 584–591.
25. Pines,G., Pines,A., Garst,A.D., Zeitoun,R.I., Lynch,S.A., Gill,R.T. *et al.* (2015) Codon compression algorithms for saturation mutagenesis. *ACS Synth. Biol.*, **4**, 604–614.
26. Halweg-Edwards,A.L., Pines,G., Winkler,J.D., Pines,A., Gill,R.T. and Web,A. (2016) Interface for codon compression. *ACS Synth. Biol.*, **5**, 1021–1023.
27. Nackley,A.G., Shabalina,S.A., Tchivileva,I.E. Satterfield,K., Korchynskyi,O., Makarov,S.S., Maixner,W., Diatchenko,L. *et al.* (2006) Human Catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science*, **314**, 1930–1933.
28. Kudla,G., Murray,A.W., Tollervey,D. and Plotkin,J.B. (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*, **324**, 255–258.
29. Komar,A.A., Lesnik,T. and Reiss,C. (1999) Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett.*, **462**, 387–391.
30. Kimchi-Sarfaty,C., Oh,J.M., Kim,I.-W., Sauna,Z.E., Calcagno,A.M., Ambudkar,S.V., Gottesman,M.M. *et al.* (2007) A 'Silent' polymorphism in the MDR 1 gene changes substrate specificity. *Science*, **315**, 525–528.