SYNTHETIC BIOLOGY

Multiplexed gene synthesis in emulsions for exploring protein functional landscapes

Calin Plesa,[1]* Angus M. Sidore,[2]* Nathan B. Lubock,[1] Di Zhang,[3] Sriram Kosuri[1,4]†

Improving our ability to construct and functionally characterize DNA sequences would broadly accelerate progress in biology. Here, we introduce DropSynth, a scalable, low-cost method to build thousands of defined gene-length constructs in a pooled (multiplexed) manner. DropSynth uses a library of barcoded beads that pull down the oligonucleotides necessary for a gene's assembly, which are then processed and assembled in water-in-oil emulsions. We used DropSynth to successfully build more than 7000 synthetic genes that encode phylogenetically diverse homologs of two essential genes in *Escherichia coli*. We tested the ability of phosphopantetheine adenylyltransferase homologs to complement a knockout *E. coli* strain in multiplex, revealing core functional motifs and reasons underlying homolog incompatibility. DropSynth coupled with multiplexed functional assays allows us to rationally explore sequence-function relationships at an unprecedented scale.

The scale at which we can build and functionally characterize DNA sequences sets the pace at which we explore and engineer biology. The recent development of multiplexed functional assays allows for the facile testing of thousands to millions of sequences for a wide array of biological functions (*1*, *2*). Currently, such assays are limited by their ability to build or access DNA sequences to test. Natural or mutagenized DNA sequences (*3*, *4*) allow for large libraries but are not easily programmed and thus limit hypotheses, applications, and engineered designs. Alternatively, researchers can use low-cost microarray-based oligo pools that allow for large libraries of designed ~200-nucleotide (nt) sequences (*5*), but their short lengths limit many other applications. Gene synthesis is capable of creating long-length sequences, but high costs currently prohibit building large libraries of designed sequences (*6*–*9*).

We developed a gene synthesis method we term DropSynth: a multiplexed approach capable of building large pooled libraries of designed gene-length sequences. DropSynth uses microarray-derived oligo libraries to assemble gene libraries at vastly reduced costs. We and others have developed robust parallel processes to build genes from oligo arrays, but because each gene must be assembled individually, costs are prohibitive for large gene libraries (*6*, *10*). In these efforts, the ability to isolate and concentrate DNA from the background pool complexity was paramount for robust assemblies (*11*). Previous efforts to multiplex such assemblies have not isolated reactions from one another and thus suffered from short assembly lengths, highly biased libraries, the inability to scale, and constraints on sequence homology (*12*–*15*).

DropSynth works by pulling down only those oligos required for a particular gene's assembly onto barcoded microbeads from a complex oligo pool. By emulsifying this mixture into picoliter droplets, we isolate and concentrate the oligos before gene assembly, overcoming the critical roadblocks for proper assembly and scalability (Fig. 1A and movie S1). The microbead barcodes are distinct 12-nt sequences that all oligos for a particular assembly share, and pair with complementary strands displayed on the microbead. Within each droplet, sequences are released from the bead by using Type IIs restriction enzyme sites and assembled through polymerase cycling assembly (PCA) into full-length genes. Last, the emulsion is broken, and the gene library is recovered. To test and optimize the protocol, we built model assemblies that were different but shared common overlap sequences. As a result, any contaminating oligo would still participate in the assembly reaction, allowing us to monitor assembly specificity and library coverage. We optimized each aspect of the protocol by trying to assemble 24-, 96-, and 288-member libraries composed of 3, 4, 5, and 6 oligos at once, based on how often we saw intended targets versus their expected frequency given random (bulk) assembly (Fig. 1B). Over many iterations, we achieved high enrichment rates (~10^8) by modifying the amount of beads, presence of size selection after assembly, ligase used for capture, and bead attachment chemistry. We ultimately found that using streptavidin bead chemistry, Taq ligase for bead capture, and size-selection after assembly yielded the highest enrichment rates. Using these protocols, we were able to build libraries of up to six oligos that produced correctly sized bands (Fig. 1C), and the resulting assembly distributions were not overly skewed (Fig. 1D and fig. S1).

To test the scalability of DropSynth, we attempted assembly of 12,672 genes ranging in size from 381 to 669 base pairs (bp) that encode homologs of two bacterial proteins from across the tree of life (Fig. 2A and fig. S2). A total of 33 libraries of 384 genes each encoded 5775 homologs of dihydrofolate reductase (DHFR) with two different codon usages (11,520 DHFR genes), as well as 1152 homologs of the enzyme phosphopantetheine adenylyltransferase (PPAT) (fig. S3, A and B). DHFR genes were assembled from either four or five 230-nt oligos, whereas PPAT genes were assembled from five 200-nt oligos. We obtained correctly sized bands for 31 of 33 assemblies, with one failing because of oligo amplification issues and the other because of low yield on the oligo processing steps, in contrast to attempts using bulk assembly that produced shorter failed by-products (fig. S3C). Three of the libraries (5x 230-nt oligomers) were too long to verify by using our barcoding approach, but the resulting synthesis showed correct band formation (fig. S4).

We cloned the libraries into an expression plasmid containing a random 20-bp barcode (assembly barcode) and sequenced the remaining 28 libraries consisting of 10,752 designs (figs. S3D, S4, and S5). For the PPAT 5x 200-nt oligo assemblies, sequencing revealed that a total of 872 genes (75%) had assemblies corresponding to a perfect amino acid sequence represented by at least one assembly barcode, with a median of two reads per assembly barcode and 56 assembly barcodes per homolog (Fig. 2B and fig. S6, A and B). This coverage increased when including sequences with deviations from the designed sequences, with 1002 genes (87%) represented within five amino acids from the designed sequences (all homologs have some alignments regardless of distance) (fig. S6D). For the DHFR 4x 230-nt oligo assemblies, we observed perfect sequences for 65% (6271) of the designed homologs, and 75% have at least one assembly within a two-amino-acid difference from design. Because there are two codon usages per homolog, when combined over homologs we observed that 3950 (79%) have at least one perfect, and 88% have at least one assembly in a distance of two amino acids (Fig. 2C). We see a strong correlation [Pearson correlation coefficient (ρ) = 0.73, P value = 3.4×10^{-5}] between the amount of DNA used to load the DropSynth beads and the resulting library coverage (fig. S7A). We also found 15 microbead barcodes that have more dropouts than would be expected by chance (fig. S7B). For constructs with at least 100 assembly barcodes, we observed a median of 1.9% (σ = 2.9%) and 3.9% (σ = 3.8%) perfect protein assemblies (Fig. 2A and figs. S6C and S8) for PPAT and DHFR libraries, respectively. The nearly double the rate of perfects for DHFR libraries compared with PPAT can be

[1]Department of Chemistry and Biochemistry, University of California, Los Angeles (UCLA), Los Angeles, CA, USA. [2]Department of Chemical and Biomolecular Engineering, UCLA, Los Angeles, CA, USA. [3]Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [4]UCLA–U.S. Department of Energy Institute for Genomics and Proteomics, Molecular Biology Institute, Quantitative and Computational Biology Institute, Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, Jonsson Comprehensive Cancer Center, UCLA, Los Angeles, CA, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: sri@ucla.edu

attributed to using longer oligos (230 versus 200 nt) that only require four oligos instead of five to assemble the gene (fig. S9A). Increasing the oligo length provides a way to assemble longer genes without large decreases in the resulting yields (fig. S9B). Furthermore, the distribution of perfect assemblies in the PPAT libraries is not overly skewed (fig. S6D), and most library members have assemblies with high identity to their respective designed homologs (fig. S6F). The resultant error profiles were consistent with Taq-derived mismatch and assembly errors that we have observed previously (fig. S10) (*16*).

We sought to show how DropSynth-assembled libraries could be easily coupled as inputs into multiplex functional assays by probing how well the PPAT homologs of various evolutionary distance to *Escherichia coli* could rescue a knockout phenotype. PPAT is an essential enzyme, encoded by the gene *coaD*, which catalyzes the second-to-last step in the biosynthesis of coenzyme A (CoA) (fig. S11) (*17*) and is an attractive target for the development of novel antibiotics (*18*). Assembled PPAT variants on the barcoded expression plasmid were transformed into *E. coli* Δ*coaD* cells and screened for complementation by growing the library in batch culture through three serial 1000-fold

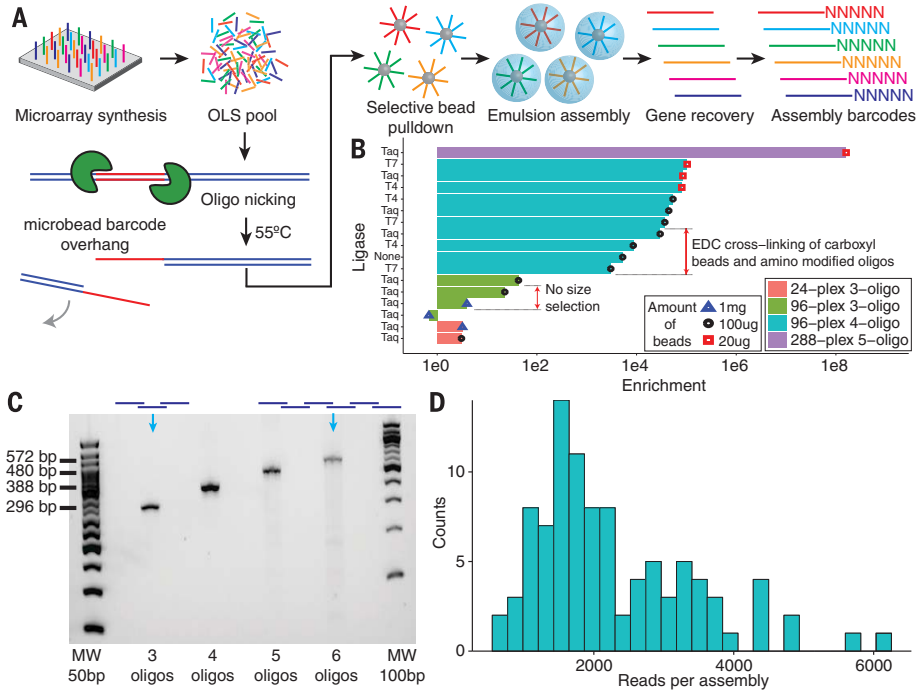

Fig. 1. DropSynth assembly and optimization. (**A**) We amplified array-derived oligos and exposed a single-stranded region that acts as a gene-specific microbead barcode. Barcoded beads display complementary single-stranded regions that selectively pull down the oligos necessary to assemble each gene. The beads are then emulsified, and the oligos are assembled by means of PCA. The emulsion is then broken, and the resultant assembled genes are barcoded and cloned. (**B**) We used a model gene library that allowed us to monitor the level of specificity and coverage of the assembly process. We then optimized various aspects of the protocol—including purification steps, DNA ligase, and bead couplings—in order to improve the specificity of the assembly reaction. Enrichment is defined as the number of specific assemblies observed relative to what would be observed by random chance in a full combinatorial assembly. (**C**) We attempted 96-plex gene assemblies with three, four, five, or six oligos, and the resultant libraries displayed the correct-sized band on an agarose gel. (**D**) The distribution of read counts for all 96 assemblies (four-oligo assembly) as determined with NGS.

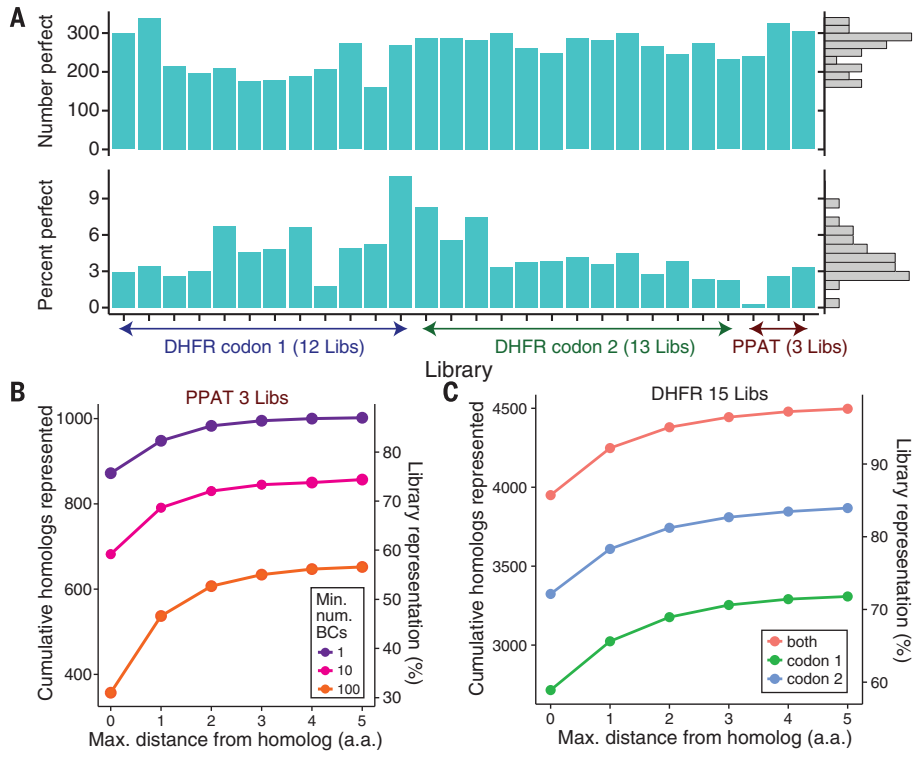

Fig. 2. DropSynth assembly of 10,752 genes. (**A**) We used DropSynth to assemble 28 libraries of 10,752 genes representing 1152 homologs of PPAT and 4992 homologs of DHFR. The number of library members with at least one perfect assembly and the median percent perfects determined by using constructs with at least 100 barcodes is shown for each library. (**B**) We observed that 872 PPAT homologs (75%) had at least one perfect assembly, and 1002 homologs (87%) had at least one assembly within a distance of five amino acids from design. (**C**) We assembled two codon variants for each designed DHFR homolog, allowing us to achieve higher coverage.

dilutions (Fig. 3A and table S1), while a rescue plasmid was simultaneously heat-cured (fig. S12). Assembly barcode sequencing of the resulting populations provided a reproducible estimate for the fitness of all homologs successfully assembled without error (biological replicates $\rho = 0.94$; Pearson, $P < 2.2 \times 10^{-16}$) (figs. S13A and S14A). Individual barcodes can display considerable noise, so having many assembly barcodes per construct improved confidence (fig. S14, B and C). Negative controls and sequences containing indels show strong depletion (figs. S13A, S15A, and S16), and fitness is reduced with increasing numbers of mutations ($\rho = -0.38$; Spearman, $P < 2.2 \times 10^{-16}$) (fig. S15, B and C). Pooled fitness scores also correlated well with measured growth rates of individually tested controls [Spearman correlation coefficient (r_s) = 0.86, $P = 5.9 \times 10^{-12}$) (fig. S17). Approximately 14% percent of the homologs show strong depletion (fitness below –2.5), whereas 70% have a positive fitness value in the pooled assay. Low-fitness homologs are evenly distributed throughout the phylogenetic tree, with only minor clustering of clades (Fig. 3B and figs. S13B, S18, and S19A). There are several reasons homologs could have low fitness, including environmental mismatches, improper folding, mismatched metabolic flux, interactions with other cytosolic components, or gene dosage toxicity effects resulting from improperly high expression (supplementary text) (*19*).

Errors during the oligo synthesis or DropSynth assembly give us mutational data across all the homologs, which we can further analyze to better understand function. We selected all 497 homologs that showed some degree of complementation (fitness greater than –1) as well as their 71,061 mapped mutants within a distance of five amino acids and carried out a multiple sequence alignment in order to find equivalent residue positions. For each amino acid and position, we found the median fitness among all of these homologs and mutants. The resulting data was projected onto the *E. coli* PPAT sequence (Fig. 4, A and B), providing data similar to deep mutational scanning approaches (*20*, *21*). We term this approach broad mutational scanning (BMS). The average BMS fitness for each position shows strong constraints in the catalytic site, at highly conserved sites ($\rho = -0.64$; Pearson, $P < 2.2 \times 10^{-16}$), and at buried residues compared with solvent-accessible ones ($\rho = 0.42$; Pearson, $P = 3.9 \times 10^{-8}$) (fig. S20, A and B, and supplementary text). Surprisingly, some residues that are known to interact with either adenosine 5′-triphosphate (ATP) or 4′-phosphopantetheine turn out to be relatively promiscuous when averaged over a large number of homologs. Furthermore, when mapped onto the *E. coli* structure (Fig. 4B), positions known to be involved with allosteric regulation by CoA or dimer formation show relatively little constraint, highlighting the diversity of distinct approaches used among different homologs while maintaining the same core function. We implemented a simple binary classifier to predict the sign of the BMS fitness value on the basis of a number of features, achieving an accuracy of 0.825 (fig. S21).

Additionally, we can search for gain-of-function (GOF) mutations among those homologs that did not complement. A total of 385 GOF mutants out of 4658 were found for 55 homologs out of 129 low-fitness homologs (fitness < –2.5). By aligning these mutations to the *E. coli* sequence, the eight statistically significant residues (34, 35, 64, 68, 69, 103, 134, and 135) shown in Fig. 4C localize to four small regions in the protein structure (fig. S22 and supplementary text). We retrieved six GOF mutants of six different homologs from the library, each with fitness determined from only a single assembly barcode, and individually tested their growth rates. Five of the six mutants showed strong growth, and one failed to complement (fig. S17B). We also tested two of the corresponding low-fitness homologs, finding increases in the growth rate of 10 and 42% for their GOF mutants (table S2).

Broad mutational scanning enabled by DropSynth is a useful tool with which to explore protein functional landscapes. By analyzing many highly divergent homologs, individual steric clashes, which might be important to a particular sequence, become averaged across the homologs. More broadly, DropSynth allows for building large designed libraries of gene-length sequences, with no specialized equipment and estimated total costs below $2 per gene (tables S3 and S4). We also show that DropSynth can be combined with dial-out polymerase chain reaction (*15*), which

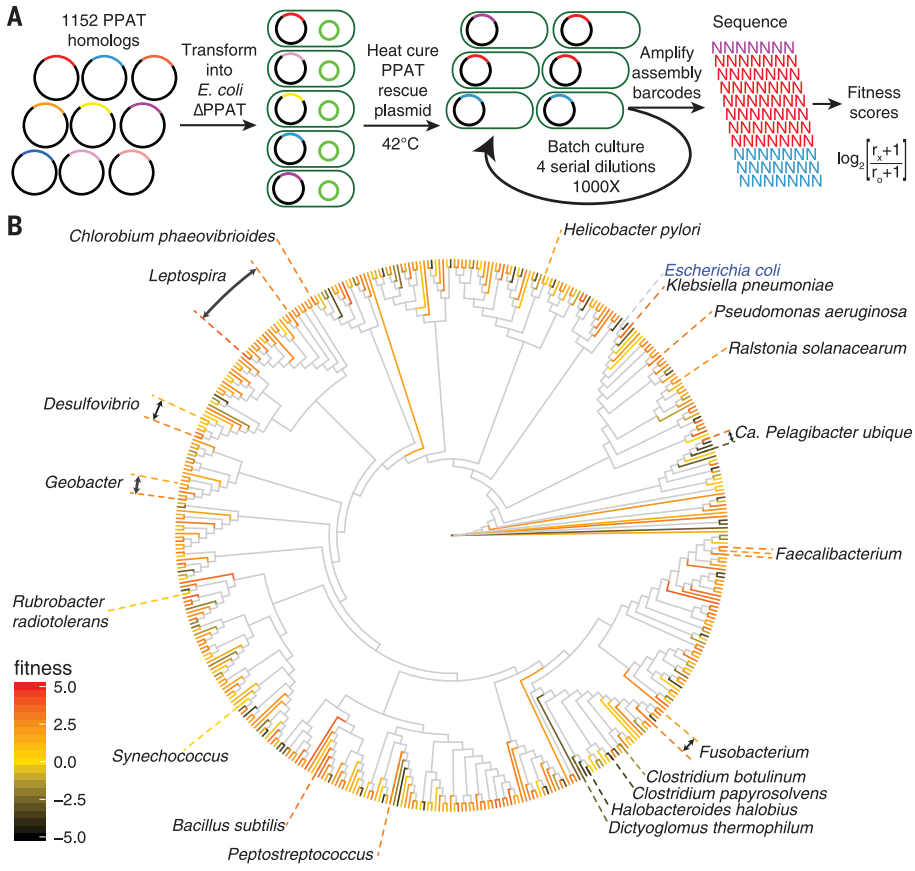

Fig. 3. PPAT complementation assay. (**A**) We used DropSynth to assemble a library of 1152 homologs of PPAT, an essential enzyme catalyzing the second-to-last step in CoA biosynthesis, and functionally characterized them using a pooled complementation assay. The barcoded library was transformed into *E. coli* Δ*coaD* cells containing a curable rescue plasmid expressing *E. coli coaD*. The rescue plasmid was removed, allowing the homologs and their mutants to compete with each other in batch culture. We tracked assembly barcode frequencies over four serial 1000-fold dilutions and used the frequency changes to assign a fitness score. (**B**) This phylogenetic tree shows 451 homologs each with at least five assembly barcodes, a subset of the full data set, in which leaves are colored by fitness. Despite having a median 50% sequence identity, we found that the majority of PPAT homologs are able to complement the function of the native *E. coli* PPAT, with 70% having positive fitness values, whereas low-fitness homologs are dispersed throughout the tree, without much clustering of clades.

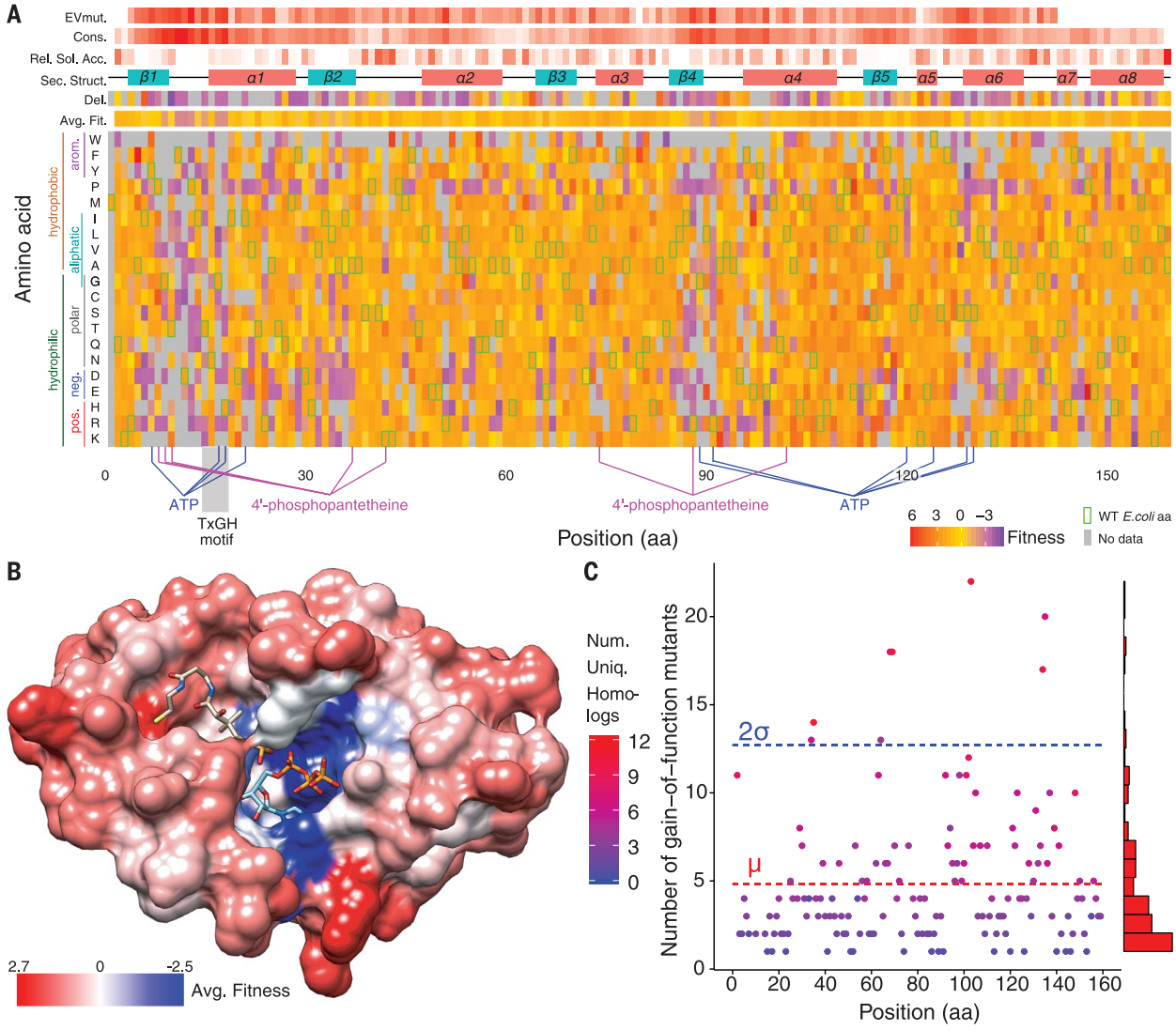

Fig. 4. Broad mutational scanning analysis. (**A**) The fitness landscape of 497 complementing PPAT homologs and their 71,061 mutants (within a distance of five amino acids) is projected onto the *E. coli* PPAT sequence, with each point in the heatmap showing the average fitness over all sequences containing that amino acid at each aligned position. Mutations are highly constrained at a core group of residues involved in catalytic function. Other positions show relatively little loss of function, when averaged over many homologs, despite known interactions with the substrates. The *E. coli* wild-type (WT) sequence is indicated by green squares, and the average position fitness, fitness of a residue deletion, mean EVmutation evolutionary statistical energy (*22*), site conservation, relative solvent accessibility, and secondary structure information is shown above. (**B**) The average fitness at each position, with blue and red representing low and high fitness, respectively, overlaid on the *E. coli* PPAT [Protein Data Bank 1QJC and 1GN8 (*23*)] structure complexed with 4′-phosphopantetheine and ATP. We observed loss of function for mutations occurring at the active site, whereas other residues involved with allosteric regulation by CoA or dimer interfaces show large promiscuity, highlighting different strategies used among homologs. (**C**) In addition to complementing homologs, we can also analyze mutants of the 129 low-fitness (<−2.5) homologs, finding 385 GOF mutants across 55 homologs. We project this data onto the *E. coli* PPAT sequence and plot the number of GOF mutants at each position, shaded by the number of different homologs represented. We found a total of eight statistically significant positions (residues 34, 35, 64, 68, 69, 103, 134, and 135) corresponding to four regions in the PPAT structure.

could be expanded for gene synthesis applications for which perfect sequences are paramount. The scale, quality, and cost of DropSynth libraries can likely be improved further with investment in algorithm design, better polymerases, and larger barcoded bead libraries.

REFERENCES AND NOTES

1. F. Inoue, N. Ahituv, *Genomics* **106**, 159–164 (2015).
2. M. Gasperini, L. Starita, J. Shendure, *Nat. Protoc.* **11**, 1782–1787 (2016).
3. K. S. Sarkisyan *et al.*, *Nature* **533**, 397–401 (2016).
4. D. M. Fowler, S. Fields, *Nat. Methods* **11**, 801–807 (2014).
5. G. J. Rocklin *et al.*, *Science* **357**, 168–175 (2017).
6. S. Kosuri, G. M. Church, *Nat. Methods* **11**, 499–507 (2014).
7. S. Ma, N. Tang, J. Tian, *Curr. Opin. Chem. Biol.* **16**, 260–267 (2012).
8. J. Quan *et al.*, *Nat. Biotechnol.* **29**, 449–452 (2011).
9. R. A. Hughes, A. D. Ellington, *Cold Spring Harb. Perspect. Biol.* **9**, a023812 (2017).
10. S. Kosuri *et al.*, *Nat. Biotechnol.* **28**, 1295–1299 (2010).
11. A. Y. Borovkov *et al.*, *Nucleic Acids Res.* **38**, e180 (2010).
12. J. C. Klein *et al.*, *Nucleic Acids Res.* **44**, e43 (2016).
13. H. Kim *et al.*, *Nucleic Acids Res.* **40**, e140 (2012).
14. T. H.-C. Hsiau *et al.*, *PLOS ONE* **10**, e0119927 (2015).
15. J. J. Schwartz, C. Lee, J. Shendure, *Nat. Methods* **9**, 913–915 (2012).
16. N. B. Lubock, D. Zhang, A. M. Sidore, G. M. Church, S. Kosuri, *Nucleic Acids Res.* **45**, 9206–9217 (2017).
17. T. Izard, A. Geerlof, *EMBO J.* **18**, 2021–2030 (1999).
18. B. L. M. de Jonge *et al.*, *Antimicrob. Agents Chemother.* **57**, 6005–6015 (2013).
19. S. Bhattacharyya *et al.*, *Eng. Life Sci.* **5**, e20309 (2016).

20. D. S. Marks, T. A. Hopf, C. Sander, *Nat. Biotechnol.* **30**, 1072–1080 (2012).
21. N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, *Cell* **138**, 774–786 (2009).
22. T. A. Hopf *et al.*, *Nat. Biotechnol.* **35**, 128–135 (2017).
23. T. Izard, *J. Mol. Biol.* **315**, 487–495 (2002).

ACKNOWLEDGMENTS

This work was supported by the funds from the Human Frontier Science Program (LT000068/2016 to C.P.), Netherlands Organisation for Scientific Research Rubicon fellowship (to C.P.), National Science Foundation Graduate Research Fellowship under grant 2016211460 (to A.M.S.), a Ruth L. Kirschstein National Research Service Award (GM007185 to N.L.), National Institutes of Health New Innovator Award (DP2GM114829 to S.K.), Searle Scholars Program (to S.K.), U.S. Department of Energy (DE-FC02-02ER63421 to S.K.), UCLA, and L. Wudl and F. Wudl. We thank J. Sampson and P. Anderson at Agilent Technologies for oligo pools and critical advice. We thank G. Church and R. Terry for guidance during the early developments and S. Feng, the UCLA Broad Stem Cell Research Center Sequencing Core, and the Technology Center for Genomics and Bioinformatics for providing next-generation sequencing (NGS) services. S.K. and D.Z. are named inventors on a patent application on the DropSynth method (US14460496). The scripts required to generate DropSynth oligos are available at https://github.com/kosurilab/DropSynth. Sequencing data are available from the sequencing read archive (SRA) with the accession no. SRP126669.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/359/6373/343/suppl/DC1
Materials and Methods
Figs. S1 to S25
Tables S1 to S14
References (*24*–*48*)
Movie S1

28 July 2017; accepted 18 December 2017
10.1126/science.aao5167

Science

Multiplexed gene synthesis in emulsions for exploring protein functional landscapes

Calin Plesa, Angus M. Sidore, Nathan B. Lubock, Di Zhang and Sriram Kosuri

Science **359** (6373), 343-347.
DOI: 10.1126/science.aao5167originally published online January 4, 2018

Large-scale gene synthesis in tiny droplets

Gene synthesis technology is important for functional characterization of DNA sequences and for the development of synthetic biology. However, current methods are limited by their low scalability and high cost. Plesa *et al.* developed a gene synthesis method, DropSynth, which uses barcoded beads to concentrate oligos and subsequently assemble them into synthetic genes within picoliter emulsion droplets. DropSynth allows generation of large libraries of thousands of genes and functional testing of all possible mutations of a particular sequence.

Science, this issue p. 343