HHS Public Access

Author manuscript

Science. Author manuscript; available in PMC 2018 November 28.

Published in final edited form as:

Science. 2018 January 19; 359(6373): 343–347. doi:10.1126/science.aao5167.

Multiplexed Gene Synthesis in Emulsions for Exploring Protein Functional Landscapes

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

[1]Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California, USA

[2]Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, Los Angeles, California, USA

[3]Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[4]UCLA-DOE 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, University of California, Los Angeles, Los Angeles, California, USA

Abstract

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 use DropSynth to successfully build >7000 synthetic genes that encode phylogenetically-diverse homologs of two essential genes in *E. 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 allow us to rationally explore sequence-function relationships at unprecedented scale.

One Sentence Summary:

A gene synthesis method, DropSynth, allows for the synthesis and characterization of thousands of genes in a pooled format.

Main Text:

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 across a

*To whom correspondence should be addressed. Tel: +1 310 825 8931; sri@ucla.edu.
†Co-first authors

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).

Here we develop 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 prior to gene assembly, thereby overcoming the critical roadblocks for proper assembly and scalability (Fig. 1A, Supplemental Movie S1). The microbead barcodes are unique 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 using Type IIs restriction enzyme sites and assembled through polymerase cycling assembly (PCA) into full length genes. Finally, the emulsion is broken and the gene library is recovered. To test and optimize the protocol, we built model assemblies that were unique, 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 (i.e. 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 type of bead chemistry, testing both EDC crosslinking of carboxyl beads and streptavidin-coupled beads. 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 6 oligos that produced correct sized bands (Fig. 1C), and the resulting assembly distributions were not overly skewed (Fig. 1D, Fig. S1).

To test the scalability of DropSynth, we attempted assembly of 12,672 genes ranging in size from 381 to 669 bp which encode homologs of two bacterial proteins from across the tree of life (Fig 2A, Fig. S2). A total of 33 libraries of 384 genes each encoded 5,775 homologs of

dihydrofolate reductase (DHFR) with two different codon usages (11,520 DHFR genes), as well as 1,152 homologs of the enzyme phosphopantetheine adenylyltransferase (PPAT) (Fig. S3, A and B). DHFR genes were assembled from either four or five 230-mer oligos while PPAT genes were assembled from five 200-mer oligos. We obtained correctly-sized bands for 31/33 assemblies, with one failing due to oligo amplification issues and the other due to low yield on the oligo processing steps, in contrast to attempts using bulk assembly which produced shorter failed by-products (Fig. S3C). Three of the libraries (5x 230-mers) were too long to verify 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 (Fig. S3D and Fig. S4, Fig. S5). For the PPAT 5x 200-mer 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 2 reads per assembly barcode and 56 assembly barcodes per homolog (Fig. 2B, Fig. S6, A and B). This coverage increased when including sequences with deviations from the designed sequences, with 1,002 genes (87%) represented within 5 aa from the designed sequences (all homologs have some alignments regardless of distance) (Fig. S6D). For the DHFR 4x 230-mer assemblies we observed perfect sequences for 65% (6,271) of the designed homologs, and 75% have at least one assembly within 2 aa difference from design. Since there are two codon usages per homolog, when combined over homologs we observe 3,950 (79%) have at least one perfect, and 88% have at least one assembly in distance 2 aa (Fig. 2C). We see a strong correlation (ρ=0.73 (Pearson), p-value=3.4E-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, Fig. S6C, Fig. S8) for PPAT and DHFR libraries respectively. The nearly double the rate of perfects for DHFR libraries compared to PPAT can be attributed to using longer oligos (230 vs. 200 nt) that only require 4 oligos instead of 5 to assemble the gene (Fig. S9A). Increasing the oligo length provides a way to assemble longer genes without significant 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(16) (Fig. S10).

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 *E. coli* could rescue a knockout phenotype. PPAT is an essential enzyme, encoded by the gene *coaD*, which catalyzes the 2nd to last step in the biosynthesis of coenzyme A (CoA)(17) (Fig. S11) 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 ΔcoaD* cells and screened for complementation by growing the library in batch culture through three serial 1000-fold dilutions (Fig. 3A, 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 ρ=0.94; Pearson, p-value <2.2E-16) (Fig. S13A, Fig. 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 (Fig. S13A, Fig. S15A, S16), and fitness is reduced with increasing numbers of mutations (ρ=−0.38; Spearman, p-value <2.2E-16) (Fig. S15, B and C). Pooled fitness scores also correlated well with measured growth rates of individually tested controls (r_s=0.86, Spearman, p-value 5.9E-12) (Fig. S17). Approximately 14% percent of the homologs show strong depletion (fitness below −2.5) while 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, Fig. S13B, Fig. S18, S19A) showing the high modularity of PPAT. 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(19) (Supplementary Text).

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 distance 5 a.a. and carried out a multiple sequence alignment 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 (ρ=−0.64; Pearson, p-value <2.2E-16), and at buried residues compared to solvent-accessible ones (ρ=0.42; Pearson, p-value 3.9E-8) (Fig. S20, A and B, Supplementary Text). Surprisingly, some residues that are known to interact with either 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 coenzyme A or dimer formation, show relatively little constraint, highlighting the diversity of distinct approaches employed among different homologs, while maintaining the same core function. We implemented a simple binary classifier to predict the sign of the BMS fitness value based on a number of features, achieving an accuracy of 0.825 (Fig. S21).

Additionally, we can search for gain-of-function (GoF) mutations amongst those homologs that did not complement. A total of 385 gain-of-function (GoF) mutants out of 4,658 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, 135) shown in Fig. 4C localize to four small regions in the protein structure (Fig. S22, 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 using DropSynth is a useful tool 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 (Table S3 & S4). We also show that DropSynth can be combined with dial-out PCR(15), which could be expanded for gene synthesis applications where 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

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 No. 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.], Department of Energy (DE-FC02-02ER63421 to S.K.), UCLA, and Linda and Fred Wudl. We thank Jeff Sampson and Paige Anderson at Agilent Technologies for oligo pools and critical advice. We thank George Church and Richard Terry for guidance during the early developments and Suhua Feng, the UCLA BSCRC Sequencing Core, and the Technology Center for Genomics & Bioinformatics for providing 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 number SRP126669.

References and Notes

1. Inoue F, Ahituv N, Decoding enhancers using massively parallel reporter assays. Genomics. 106, 159–164 (2015). [PubMed: 26072433]
2. Gasperini M, Starita L, Shendure J, The power of multiplexed functional analysis of genetic variants. Nat. Protoc 11, 1782–1787 (2016). [PubMed: 27583640]
3. Sarkisyan KS et al., Local fitness landscape of the green fluorescent protein. Nature. 533, 397–401 (2016). [PubMed: 27193686]
4. Fowler DM, Fields S, Deep mutational scanning: a new style of protein science. Nat. Methods 11, 801–807 (2014). [PubMed: 25075907]
5. Rocklin GJ et al., Global analysis of protein folding using massively parallel design, synthesis, and testing. Science. 357, 168–175 (2017). [PubMed: 28706065]
6. Kosuri S, Church GM, Large-scale de novo DNA synthesis: technologies and applications. Nat. Methods 11, 499–507 (2014). [PubMed: 24781323]
7. Ma S, Tang N, Tian J, DNA synthesis, assembly and applications in synthetic biology. Curr. Opin. Chem. Biol 16, 260–267 (2012). [PubMed: 22633067]
8. Quan J et al., Parallel on-chip gene synthesis and application to optimization of protein expression. Nat. Biotechnol 29, 449–452 (2011). [PubMed: 21516083]
9. Hughes RA, Ellington AD, Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. Cold Spring Harb. Perspect. Biol 9 (2017), doi:10.1101/cshperspect.a023812.
10. Kosuri S et al., Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. Nat. Biotechnol 28, 1295–1299 (2010). [PubMed: 21113165]

11. Borovkov AY et al., High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides. Nucleic Acids Res. 38, e180 (2010). [PubMed: 20693531]
12. Klein JC et al., Multiplex pairwise assembly of array-derived DNA oligonucleotides. Nucleic Acids Res. 44, e43 (2016). [PubMed: 26553805]
13. Kim H et al., "Shotgun DNA synthesis" for the high-throughput construction of large DNA molecules. Nucleic Acids Res. 40, e140 (2012). [PubMed: 22705793]
14. Hsiau TH-C et al., A method for multiplex gene synthesis employing error correction based on expression. PLoS One. 10, e0119927 (2015). [PubMed: 25790188]
15. Schwartz JJ, Lee C, Shendure J, Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules. Nat. Methods 9, 913–915 (2012). [PubMed: 22886093]
16. Lubock NB, Zhang D, Church GM, Kosuri S, A systematic comparison of error correction enzymes by next-generation sequencing. bioRxiv (2017), p. 100685.
17. Izard T, Geerlof A, The crystal structure of a novel bacterial adenylyltransferase reveals half of sites reactivity. EMBO J. 18, 2021–2030 (1999). [PubMed: 10205156]
18. de Jonge BLM et al., Discovery of inhibitors of 4'-phosphopantetheine adenylyltransferase (PPAT) to validate PPAT as a target for antibacterial therapy. Antimicrob. Agents Chemother 57, 6005–6015 (2013). [PubMed: 24041904]
19. Bhattacharyya S et al., Transient protein-protein interactions perturb E. coli metabolome and cause gene dosage toxicity. eLife Sciences. 5, e20309 (2016).
20. Marks DS, Hopf TA, Sander C, Protein structure prediction from sequence variation. Nat. Biotechnol 30, 1072–1080 (2012). [PubMed: 23138306]
21. Halabi N, Rivoire O, Leibler S, Ranganathan R, Protein sectors: evolutionary units of three-dimensional structure. Cell. 138, 774–786 (2009). [PubMed: 19703402]
22. Hopf TA et al., Mutation effects predicted from sequence co-variation. Nat. Biotechnol 35, 128–135 (2017). [PubMed: 28092658]
23. Izard T, The crystal structures of phosphopantetheine adenylyltransferase with bound substrates reveal the enzyme's catalytic mechanism. J. Mol. Biol 315, 487–495 (2002). [PubMed: 11812124]
24. Notredame C, Higgins DG, Heringa J, T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol 302, 205–217 (2000). [PubMed: 10964570]
25. Stamatakis A, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30, 1312–1313 (2014). [PubMed: 24451623]
26. Finn RD et al., InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. 45, D190–D199 (2017). [PubMed: 27899635]
27. Eroshenko N, Kosuri S, Marblestone AH, Conway N, Church GM, in Current Protocols in Chemical Biology (John Wiley & Sons, Inc., 2009).
28. Buschmann T, Bystrykh LV, Levenshtein error-correcting barcodes for multiplexed DNA sequencing. BMC Bioinformatics. 14, 272 (2013). [PubMed: 24021088]
29. Kent WJ, BLAT—The BLAST-Like Alignment Tool. Genome Res. 12, 656–664 (2002). [PubMed: 11932250]
30. Kuhlman TE, Cox EC, Site-specific chromosomal integration of large synthetic constructs. Nucleic Acids Res. 38, e92 (2010). [PubMed: 20047970]
31. Lou C, Stanton B, Chen Y-J, Munsky B, Voigt CA, Ribozyme-based insulator parts buffer synthetic circuits from genetic context. Nat. Biotechnol 30, 1137–1142 (2012). [PubMed: 23034349]
32. Pédelacq J-D, Cabantous S, Tran T, Terwilliger TC, Waldo GS, Engineering and characterization of a superfolder green fluorescent protein. Nat. Biotechnol 24, 79–88 (2006). [PubMed: 16369541]
33. Datsenko KA, Wanner BL, One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc. Natl. Acad. Sci. U. S. A 97, 6640–6645 (2000). [PubMed: 10829079]
34. Cox RS, 3rd, Dunlop MJ, Elowitz MB, A synthetic three-color scaffold for monitoring genetic regulation and noise. J. Biol. Eng 4, 10 (2010). [PubMed: 20646328]
35. Zorita E, Cuscó P, Filion GJ, Starcode: sequence clustering based on all-pairs search. Bioinformatics. 31, 1913–1919 (2015). [PubMed: 25638815]

36. Saito T, Rehmsmeier M, Precrec: fast and accurate precision-recall and ROC curve calculations in R. Bioinformatics. 33, 145–147 (2017). [PubMed: 27591081]

37. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y, ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol 8, 28–36 (2017).

38. Pettersen EF et al., UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem 25, 1605–1612 (2004). [PubMed: 15264254]

39. Capra JA, Singh M, Predicting functionally important residues from sequence conservation. Bioinformatics. 23, 1875–1882 (2007). [PubMed: 17519246]

40. Kabsch W, Sander C, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22, 2577–2637 (1983). [PubMed: 6667333]

41. Bloom JD, An experimentally determined evolutionary model dramatically improves phylogenetic fit. Mol. Biol. Evol 31, 1956–1978 (2014). [PubMed: 24859245]

42. Izard T, A novel adenylate binding site confers phosphopantetheine adenylyltransferase interactions with coenzyme A. J. Bacteriol 185, 4074–4080 (2003). [PubMed: 12837781]

43. Hu P et al., Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. PLoS Biol. 7, e96 (2009). [PubMed: 19402753]

44. Butland G et al., Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature. 433, 531–537 (2005). [PubMed: 15690043]

45. Freiberg C et al., Identification of novel essential Escherichia coli genes conserved among pathogenic bacteria. J. Mol. Microbiol. Biotechnol 3, 483–489 (2001). [PubMed: 11361082]

46. Baba T et al., Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol 2, 2006.0008 (2006).

47. Gerdes SY et al., From genetic footprinting to antimicrobial drug targets: examples in cofactor biosynthetic pathways. J. Bacteriol 184, 4555–4572 (2002). [PubMed: 12142426]

48. Geerlof A, Lewendon A, Shaw WV, Purification and characterization of phosphopantetheine adenylyltransferase from Escherichia coli. J. Biol. Chem 274, 27105–27111 (1999). [PubMed: 10480925]

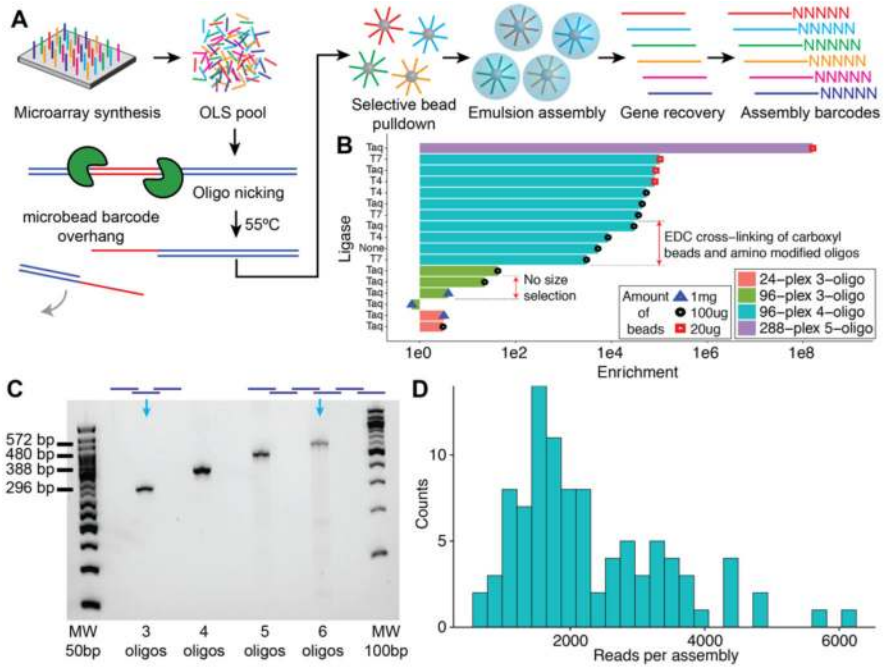

Figure 1. DropSynth assembly and optimization.

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 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 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 3, 4, 5, or 6 oligonucleotides and the resultant libraries displayed the correct-sized band on an agarose gel. **D)** The distribution of read-counts for all 96 assemblies (4-oligo assembly) as determined by NGS.

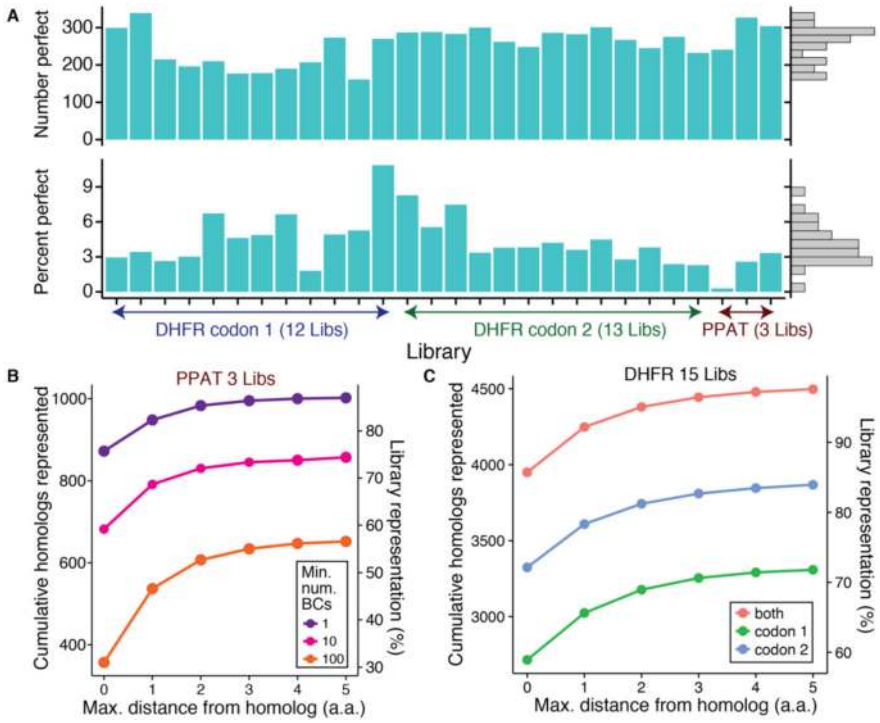

Figure 2. DropSynth assembly of 10,752 genes.

A) We used DropSynth to assemble 28 libraries of 10,752 genes representing 1,152 homologs of PPAT and 4,992 homologs of DHFR. The number of library members with at least one perfect assembly and the median percent perfects determined using constructs with at least 100 barcodes is shown for each library. **B)** We observe that 872 PPAT homologs (75%) had at least one perfect assembly, and 1,002 homologs (87%) had at least one assembly within a distance of 5 a.a. from design. **C)** We assembled two codon variants for each designed DHFR homolog, allowing us to achieve higher coverage.

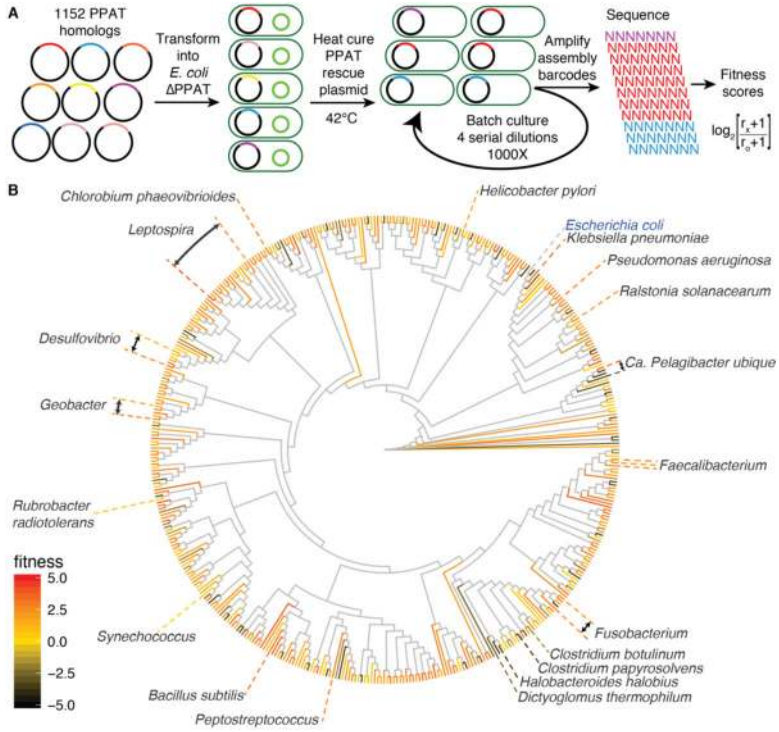

Figure 3. PPAT complementation assay.

A) We used DropSynth to assemble a library of 1152 homologs of phosphopantetheine adenylyltransferase (PPAT), an essential enzyme catalyzing the second-to-last step in coenzyme A biosynthesis, and functionally characterized them using a pooled complementation assay. The barcoded library was transformed into *E. coli ΔcoaD* 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 a 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 5 assembly barcodes, a subset of the full data set, where leaves are colored by fitness. Despite having a median 50% sequence identity, we find that the majority of PPAT homologs are able to complement the function of the native *E. coli* PPAT, with 70% having positive fitness values, while low-fitness homologs are dispersed throughout the tree without much clustering of clades.

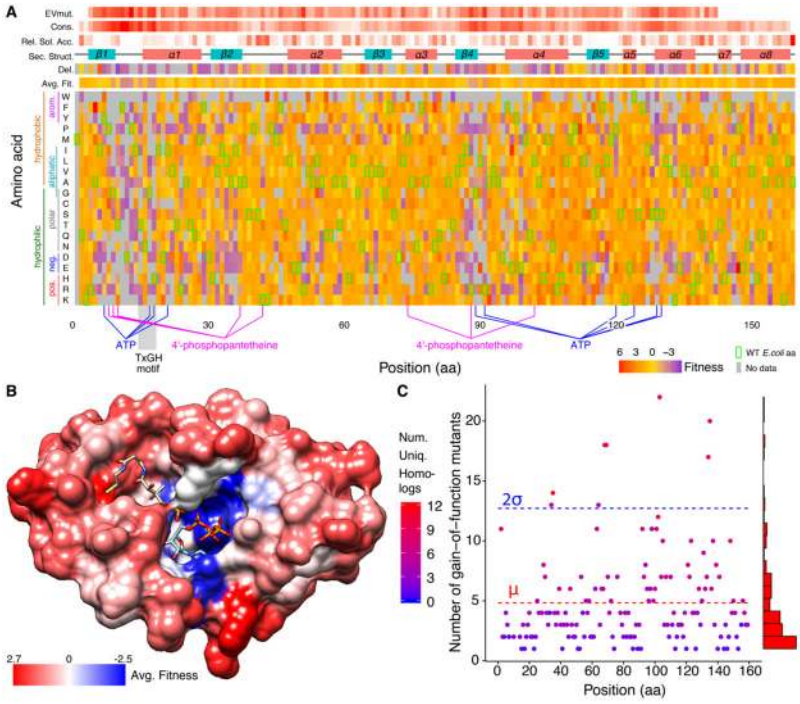

Figure 4. Broad mutational scanning (BMS) analysis

A) The fitness landscape of 497 complementing PPAT homologs and their 71,061 mutants (within a distance of 5 a.a.) 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* WT sequence is indicated by green squares, while 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 (PDB: 1QJC, 1GN8(23)) structure complexed with 4'-phosphopantetheine and ATP. We observe loss-of-function for mutations occurring at the active site, while other residues involved with allosteric regulation by coenzyme A or dimer interfaces show large promiscuity, highlighting different strategies employed among homologs. **C)** In addition to complementing homologs, we can also analyze mutants of the 129 low-fitness (< −2.5) homologs, finding 385 gain-of-function (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 find a total of 8 statistically significant positions (residues: 34, 35, 64, 68, 69, 103, 134, 135) corresponding to four regions in the PPAT structure.