Multiplexed gene synthesis in emulsions for exploring protein functional landscapes

Calin Plesa,1* Angus M. Sidore,2* Nathan B. Lubock,1 Di Zhang,3 Siriram Kosuri1,4†

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 adenyllyltransferase 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 covered. To test and optimize the protocol, we built model assemblies that were different but covered. To test and optimize the protocol, we built model assemblies that were different but covered.
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.
creasing numbers of mutations (S15A, and S16), and fitness is reduced with in- 
taining indels show strong depletion (figs. S13A, codes per construct improved confidence (fig. S14, siderable noise, so having many assembly bar-
individual barcodes can display con-
E. coli coaD rescue plasmid expressing = 5.9 × 10
control [Spearman correlation coefficient (r
high expression (supplementary text) (sup-
actions with other cytosolic components, or gene
There are several reasons homologs could have
phylogenetic tree, with only minor clustering of
percent of the homologs show strong depletion
P
< 2.2 × 10−16) (fig. S15, B and C).
Pooled fitness scores also correlated well with
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 homo-
logs that showed some degree of complementa-
tion (fitness greater than −1) as well as their
71,061 mapped mutants within a distance of five
amino acids and carried out a multiple se-
quence alignment in order to find equivalent
residue positions. For each amino acid and posi-
tion, 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 posi-
tion shows strong constraints in the catalytic site,
at highly conserved sites (r = −0.64; Pearson,
P < 2.2 × 10−16), and at buried residues compared
with solvent-accessible ones (r = 0.42; Pearson,
P = 3.9 × 10−6) (fig. S20, A and B, and supplemen-
tary 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 aver-
aged over a large number of homologs. Further-
more, 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 dif-
erent 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 align-
ing 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 se-
cence, 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 (35), 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 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.
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


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.
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 (GM037135 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 developmenents 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
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

ARTICLE TOOLS http://science.sciencemag.org/content/359/6373/343
SUPPLEMENTARY MATERIALS http://science.sciencemag.org/content/suppl/2018/01/03/science.aao5167.DC1
REFERENCES This article cites 47 articles, 9 of which you can access for free http://science.sciencemag.org/content/359/6373/343#BIBL
PERMISSIONS http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service