Published on Dec 02, 2023 by Arcadia Science **DOI**: 10.57844/arcadia-daf9-5609

Combinatorial indexing and screening of clonal DNA fragments

Oligo pools can contain millions of unique sequences, but they're limited by length, error rate, and bias. We propose methods to scalably screen synthetic DNA libraries, so an individual researcher can obtain thousands of error-free synthetic DNA assemblies at low cost.

Contributors (A-Z)

Prachee Avasthi, Jase Gehring, Megan L. Hochstrasser, Ilya Kolb

Version 2 · Mar 31, 2025

Purpose

We present ideas for methods that researchers can use to screen thousands of clonal DNA fragments using combinatorial indexing, pooled library preparation, and long-read DNA sequencing. Using oligo pools as a source of synthetic DNA, we show how individual researchers can, in principle, obtain thousands of synthetic genes at a cost as low as $1 per kilobase of synthetic DNA.

While Arcadia isn’t pursuing these ideas and we haven’t tested them, we think they might be useful to other scientists interested in using high-quality, inexpensive DNA arrays.

We’ve put this effort on ice! ☒

#TranslationalMismatch

We're not positioned to have a strategic advantage in this space. While we're intrigued by many of these ideas and hope others develop them further, this isn't an area Arcadia currently plans to pursue commercially.

Learn more about the Icebox and the different reasons we ice projects.

Motivation

As of Fall 2023, a typical synthetic gene that is 1,000 base pairs (bp) long might cost $70 USD from commercial suppliers like IDT or Twist Bioscience. At this rate of$ 0.07 per bp of DNA, experiments using synthetic genes are well within the reach of most laboratories. But, despite drastic reductions in the cost of synthetic DNA, very large-scale projects are still limited by the cost and capacity of DNA synthesis. Routine access to large-scale gene synthesis (> 1 Mb of DNA) will unlock new experiments in synthetic biology, such as evaluating synthetic chromosomes and designing protein libraries.

This pub shows how two breakthrough technologies — DNA oligo pool synthesis and long-read DNA sequencing — could allow researchers to prepare synthetic arrays of thousands of genes at greatly reduced cost compared to commercial gene synthesis. Oligo pools offer up to 100× reduction in cost per base pair, but the complex pools must be assembled and arrayed for DNA assembly projects. Our parallel colony sequencing concept isolates and identifies correct DNA constructs. We suggest this

approach to researchers who require synthetic DNA arrays for large-format biochemical and genetic experiments.

The idea

DNA synthesis drives progress in biotechnology, and recent breakthroughs in parallel and miniaturized DNA synthesis have opened up entirely new possibilities for large-scale experiments **[1]**. Today, electrochemical DNA synthesis can yield oligo pools with thousands of unique DNA fragments in a time- and cost-effective manner. Companies like Twist Bioscience offer DNA oligo pools up to 300 nucleotides (nt) in length and with practically unlimited pool size. But there are still serious challenges for researchers to fully utilize this powerful synthetic platform. The synthesized oligos are shorter than most genes, requiring gene assembly, and error rates and library bias hamper the use of oligo pools in a range of applications from CRISPR screens to parallel mutagenesis **[2]**.

Scalable methods for sequence verification would allow researchers to sift through oligo pool DNA and obtain clonal, sequence-verified, and balanced libraries needed to realize the full potential of complex DNA oligo pools. One such approach, uPIC-M **[3]**, was demonstrated in 2021 and involves indexed colony PCR of gene libraries cloned into bacterial colonies. Another has recently been demonstrated using bacterial conjugation for parallel plasmid barcoding followed by pooled library preparation and long-read sequencing **[4]**. Crucially, long-read sequencing technology has recently achieved sufficient accuracy and scale to support such an application **[5]**.

Here we present additional possibilities for indexed screening of bacterial colonies containing clonal DNA fragments. Our workflows largely extend the uPIC-M method mentioned above, with the addition of combinatorial indexing steps and the incorporation of Nanopore sequencing for full-length gene sequence verification. These modifications would greatly increase the capacity of the uPIC-M method and decrease the cost of each correct clone. It should be feasible to screen tens of thousands of DNA fragments, yielding thousands of clonally verified gene assemblies at greatly reduced cost.

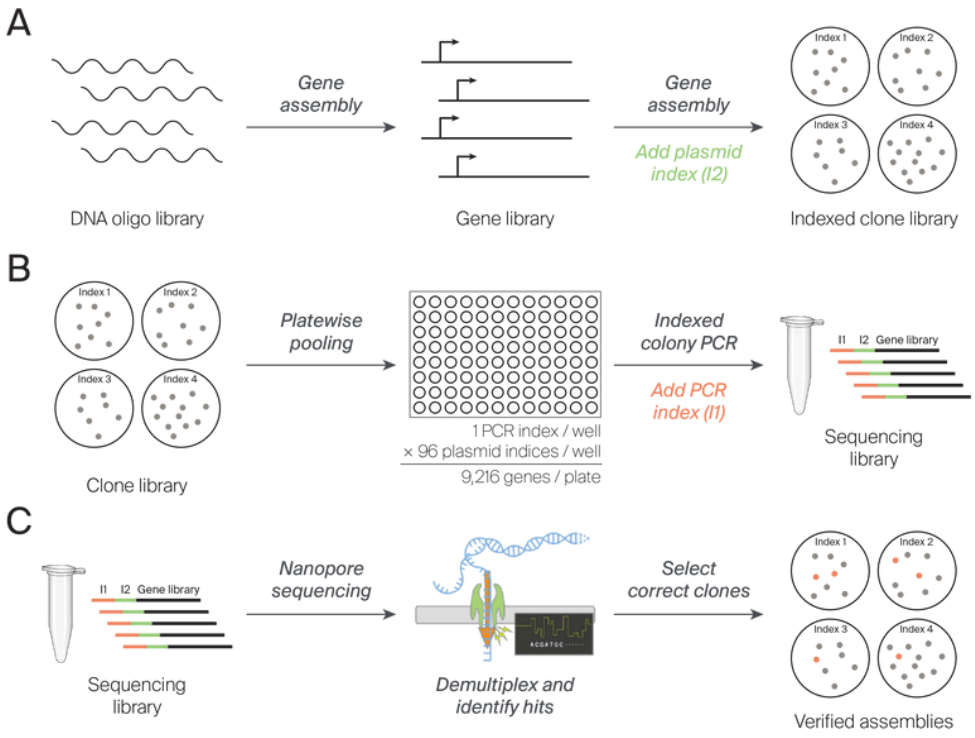

Figure 1

Combinatorial indexing for parallel gene synthesis.

(A) Gene libraries can be obtained from commercial suppliers or by custom gene synthesis. The gene library is cloned into *E. coli* to physically isolate and clonally amplify the template DNA. Parallel transformation using indexed plasmids provides a first layer of indexing.

(B) Colonies are picked into 96-well plates, pooling one colony from each indexed transformation into each well of the plate. Each pool is amplified with well-specific, indexed PCR primers, resulting in a library with PCR indices ("I1" above) and plate indices ("I2" above).

(C) The indexed library is pooled, prepared for sequencing, and analyzed to identify 100% correct synthetic genes and the colony associated with each gene. Correct clones can be re-pooled or arrayed for downstream use.

NOTE: We haven't tested these methods but hope readers will, so in the following sections, we use imperative tense to describe what one might do to implement our approach.

Initial library preparation

First, follow existing methods for library manipulation to prepare a plasmid library containing synthetic DNA clones**[6][7][8]**. You may choose to perform error correction or DNA assembly steps at this stage to reduce the screening burden. Other experiments might involve gene shuffling or mutagenesis, as in the uPIC-M manuscript, or simply arraying and normalizing the purchased oligo pool. Regardless of the library preparation method, once you obtain the desired library, clone it into a bacterial plasmid backbone, with the specific plasmid design described below. In most implementations, transform libraries into a panel of indexed destination plasmids as a primary indexing step (Figure 1, A).

Indexed destination plasmids and pooled colony PCR

In a first realization, you might use combinatorial indexing to uniquely associate a bacterial colony with a two-part index (Figure 1). You install the first index by cloning into the destination plasmid, and add a second index via pooled colony PCR. Here's an example: Prepare a set of 96 destination plasmids with plasmid-specific indices ("plasmid indices"). Introduce a library of DNA fragments into each indexed destination plasmid, and plate the resulting colonies onto separate plates for each plasmid index (Figure 1, A). Now, you can combine one colony from each plate into each well of a 96-well plate. Grow the colonies as pools — one plate can accommodate 96 colonies × 96 wells = 9,216 clones (Figure 1, B). You can use indexed 96-well colony PCR to add additional well-specific indices, after which you can combine and process all wells as a pool for library preparation and long-read sequencing. Accurate, high-depth long-read sequencing covers both the indices and the DNA fragment, revealing 100% correct DNA fragments and the location of the originating colony (Figure 1, C). This two-level barcoding scheme should be capable of processing 9,216 colonies at a cost of around

$1,000, including Nanopore sequencing and indexed colony PCR (Table 1). This cost estimate is in the range of that estimated by Li et al. **[4]**. Assuming you need to sample seven colonies per gene, as in the uPIC-M method, you can use a single PCR plate to synthesize > 1,300 unique genes at an estimated cost of < $2 per gene (Table 1). Similar schema may involve 384- or 1,536-well plates or different numbers of plasmids with plasmid indices.

The main drawback of this approach is the reliance on PCR amplification, which is likely to lead to bias in the library composition and require deeper sequencing to recover all input DNAs. PCR amplification will also introduce additional mutations in the DNA sequence that may increase the sequencing depth needed to identify correct clones. Finally, performing parallel processing steps (96× bacterial transformations, 96× PCR across an entire plate) adds cost and complexity to the overall workflow, albeit with a substantial increase in the screening capacity.

Number of 1 kb genes	**20.4**	**204**	**2,040**	**20,400**	**20 gene fragments**
Number of oligos	120	1,200	12,000	120,000	20 blocks
Oligo length (bp)	170 bp	170 bp	170 bp	170 bp	1,000 bp
Oligo pool cost	$598.50	$2,400	$2,420	$9,000	$1,400
Supplier	IDT	GenScript	GenScript	GenScript	IDT
In vitro assembly scale	1× tube	10× tubes	1× 96-well plate	10× 96-well plate	NA
In vitro assembly cost	$5	$50	$500	$5,000	NA
Transformation cost	$1	$10	$100	$1,000	$1
Colonies screened	140	1,400	14,000	140,000	140
100× long-read sequencing	2.04 Mb	20 Mb	200 Mb	2 Gb	2.04 Mb
Sequencing cost	$0.30	$3	$30	$300	$0.30
Total cost	$607.8	$2,463	$3,050	$15,300	$1,401
Cost per perfect gene	$30.35	$12.07	$1.50	$0.75	$70

Table 1

Hypothetical synthesis projects are shown for increasing numbers of genes to be synthesized. Pricing for oligo pool DNA substrate as provided by the supplier is shown. Prices for *in vitro* assembly, transformation, and sequencing are estimates based on typical pricing for these steps and will depend on pricing and quantities required in specific embodiments of these ideas. Note that we've left labor costs out of this estimate, and roughly estimated the cost of *in vitro* assembly. Substantial cost reductions are achievable compared to using pre-

prepared gene fragments (right-most column). At a scale of thousands of genes, it is feasible for a synthetic gene to cost less than $1 in consumables expenses.

Increased scale with additional levels of indexing

We present alternative ways to introduce indices to clonal genes. These strategies are more complex than the general approach outlined above, but they potentially allow for larger libraries to be screened, further reducing the cost per gene.

Adding more indices during library preparation

You could pool multiple plates of PCR amplicons by adding a third level of indexing. Most simply, use a second round of indexed PCR amplification to pool multiple plates and increase the number of cloned fragments captured in a single sequencing library. Or, add an index by ligation or by tagmenting plasmid DNA with DNA-barcoded Tn5 transposase complexes prior to pooling and PCR amplifying. It's likely that colony picking and sequencing depth will be more limiting in practice than the number of indices that you can effectively multiplex.

Adding more indices during plasmid assembly

A simple and effective way to introduce additional plasmid-level indices is by incorporating a short index as a DNA fragment during the plasmid assembly reaction. For example, if you use a library of pre-indexed plasmids for cloning, adding another combinatorial plasmid index as a DNA fragment can multiply the number of unique plasmids used without requiring explicit cloning and maintenance of a larger number of plasmids. Again, plating each unique plasmid transformation on its own plate and tracking the colony from the plate to a specific well on the PCR plate ensures unique identification of the originating colony by the indexed DNA sequence.

Fluorescent colony indexing

You could achieve an alternative or additional layer of plasmid indexing by fluorescent cell barcoding **[9]**. Here, design plasmids with multiple fluorescent reporters such that the combination of fluorescent signals produces a unique fluorescence spectrum specific to each plasmid. This kind of fluorescent index can replace or act in addition to the plasmid index described above. Because you can analyze fluorescent spectra non-invasively on living colonies, these indices let you identify positive colonies after sequencing based on their fluorescence spectrum alone. In most embodiments, you'd use colony isolation to isolate and amplify clonal fragments before library preparation, meaning fluorescent indexing is not particularly advantageous, but fluorescent indexing is unique in that you can identify cells and colonies by fluorescence *after* multiplexed sequencing. There may well be certain embodiments or applications where fluorescence indexing can be highly enabling.

Realizing this approach would require first associating fluorescence spectra with a plasmid library. The plasmids would encode unique fluorescence signatures through a combination of various promoters and fluorescent proteins. Theoretically, three fluorescent proteins with five different promoters would give $5^3 = 125$ combinations of fluorescence intensity that you could uniquely identify.

Spatial indexing

You can transfer spatial arrays of DNA indices to bacterial colonies in a fashion analogous to modern spatial genomics experiments **[10]**. In an example embodiment, replica-plate bacterial colonies containing plasmids with synthetic DNA inserts with the donor plate used for downstream growth and the recipient plate used for indexing and sequencing. Fix the recipient plate of colonies *in situ* within a polymeric hydrogel, such as a polyacrylamide gel, and then assign indices via primer extension or PCR using an array of indexing oligos. You can then pool the indexing DNAs, potentially with an additional round of PCR-based indexing, and prepare a pooled library for long-read sequencing.

Summary

DNA oligo pools have upended the cost of synthetic DNA. With millions of unique oligos prepared in parallel on a single chip, the price per base pair has plummeted

dramatically. Using oligo pool DNA in experiments, however, comes with its own challenges. Oligo pools must be amplified, manipulated, and characterized. They suffer from short length, high error rate, and representation bias that prohibit advanced experiments. Our approach provides for highly parallelized cloning of synthetic DNA. With the ability to screen > 104 plasmids, you should be able to routinely prepare large arrays of synthetic DNA at a cost well below commercial prices.

Weigh in!

While these ideas don't make sense for our company to pursue at the moment, we hope others will. If you have questions or other thoughts on this pub, please leave a comment! We'd also love to hear how it goes if you try any of these approaches.

References

1. Hoose A, Vellacott R, Storch M, Freemont PS, Ryadnov MG. (2023). DNA synthesis technologies to close the gene writing gap. https://doi.org/10.1038/s41570-022-00456-9

2. Kuiper BP, Prins RC, Billerbeck S. (2021). Oligo Pools as an Affordable Source of Synthetic DNA for Cost-Effective Library Construction in Protein- and Metabolic Pathway Engineering. https://doi.org/10.1002/cbic.202100507

3. Appel MJ, Longwell SA, Morri M, Neff N, Herschlag D, Fordyce PM. (2021). uPIC–M: Efficient and Scalable Preparation of Clonal Single Mutant Libraries for High-Throughput Protein Biochemistry. https://doi.org/10.1021/acsomega.1c04180

4. Li W, Miller D, Liu X, Tosi L, Chkaiban L, Mei H, Hung P-H, Parekkadan B, Sherlock G, Levy SF. (2023). Arrayed*in vivo*barcoding for multiplexed sequence verification of plasmid DNA and demultiplexing of pooled libraries. https://doi.org/10.1101/2023.10.13.562064

5. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, Albertsen M. (2022). Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and

metagenomes without short-read or reference polishing. https://doi.org/10.1038/s41592-022-01539-7

6 Plesa C, Sidore AM, Lubock NB, Zhang D, Kosuri S. (2018). Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. https://doi.org/10.1126/science.aao5167

7 Lund S, Potapov V, Johnson SR, Buss J, Tanner NA. (2023). Highly parallelized construction of DNA from low-cost oligonucleotide mixtures using Data-optimized Assembly Design and Golden Gate. https://doi.org/10.1101/2023.11.20.567888

8 Wan W, Lu M, Wang D, Gao X, Hong J. (2017). High-fidelity de novo synthesis of pathways using microchip-synthesized oligonucleotides and general molecular biology equipment. https://doi.org/10.1038/s41598-017-06428-0

9 Krutzik PO, Nolan GP. (2006). Fluorescent cell barcoding in flow cytometry allows high-throughput drug screening and signaling profiling. https://doi.org/10.1038/nmeth872

10 Srivatsan SR, Regier MC, Barkan E, Franks JM, Packer JS, Grosjean P, Duran M, Saxton S, Ladd JJ, Spielmann M, Lois C, Lampe PD, Shendure J, Stevens KR, Trapnell C. (2021). Embryo-scale, single-cell spatial transcriptomics. https://doi.org/10.1126/science.abb9536