MULTIPLEXED DNA ASSEMBLY FROM OLIGO-POOLS

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, DNA synthesis technologies has served as a powerful tool to validate hypothesis, greatly advanced many research areas. From simpliy validating the functionality of AI-generated genetic circuitus to deepening our understanding on life's basic principles by synthesizing a genome. However, the cost of synthesizing gene-sized DNA sequences at scale remains too high, making it economically unfeasible to validate the experimental performance of ML models or generate new datasets for futher training. To address this challenge, we developed a novel method for the high-throughput assembly of gene-sized DNA sequences, starting from cost-effective chip-synthesized oligo-pools. Unlike traditional Polymerase Cycling Assembly (PCA) methods, we employed Golden Gate Assembly (GGA) to facilitate the ligation of short DNA fragments. This approach enabled us to successfully assemble high-quality DNA libraries containing up to 96 gene-sized sequences (~600 bp) in a single-pot reaction, with convenient retrieval of individual sequences. Leveraging automation to conduct numerous reactions in parallel, we can efficiently and economically synthesize gene-sized DNA sequences at scale, thereby facilitating the experimental data generation for the ML community.

1 INTRODUCTION

In recent years, DNA synthesis technologies has served as a powerful tool to validate hyperthesis, greatly advanced many research areas. From simpliy validating the functionality of AI-generated genetic circuitus to deepening our understanding on life's basic principles by synthesizing a genome. However, the cost of synthesizing gene-sized DNA sequences at scale remains too high, making it economically unfeasible to validate the experimental performance of ML models or generate new datasets for futher training.

To synthesize a gene-sized DNA sequence, it is first split into short oligonucleotides with complementary overhangs. These oligonucleotides are then separately synthesized in columns. After annealing and gap filling by polymerases, the target sequence is assembled from these short oligonucleotides and verified by clone sequencing. This process, termed Polymerase Cycling Assembly (PCA), is the main practical option for making gene-sized DNA sequences (Hoose et al., 2023).

Therefore, the cost of DNA synthesis primarily comes from three aspects: oligos, assembly cost, and error-free sequence retrieval by clone sequencing. Over the past two decades, substantial efforts have been made to reduce the cost of DNA synthesis in these areas. For example, cost-effective chip-synthesized oligo-pools have demonstrated their utility in replacing column-synthesized oligos (Tian et al., 2004; Kosuri et al., 2010; Quan et al., 2011; Plesa et al., 2018) ; another approach is to assemble DNA sequences in a library format rather than individually, thereby reducing the average assembly cost (Klein et al., 2016; Moravec et al., 2024); the application of error correction enzymes and dial-out PCR after high-throughput sequencing have also decreased the error-rate, facilitating with clone sequencing (Kosuri et al., 2010; Saaem et al., 2012; Sequeira et al., 2016). However, to the best of our knowledge, none of these methods produces large DNA libraries long enough ($> 500bp$) while maintaining a high error-free ratio ($> 50\%$) and ensuring convenient retrieval without relying on high-throughput sequencing.

In this paper, we present the following contributions:

1. We propose a novel method for precisely assembling up to 96 pairs of short DNA fragments into longer sequences in a single-pot reaction.

2. We achieved the two-step hierarchical assembly of four fragments by introducing an alternative Type IIS restriction enzyme, *BsmBI*, and demonstrated its feasibility for achieving a convenient multi-step hierarchical assembly by blocking *BsaI* digestion through methylation.

3. Starting from cost-effective chip-synthesized oligo-pools, we successfully assembled high-quality DNA libraries with up to 96 gene-sized sequences (~600 bp) in a single-pot reaction and retrieved individual sequences conveniently. Leveraging automation to conduct numerous reactions in parallel, we can efficiently and economically synthesize gene-sized DNA sequences at scale.

2 METHODS:

In contrast to the traditional PCA method, we employed Golden Gate Assembly (GGA) to facilitate the ligation of short DNA fragments. GGA leverages the unique properties of Type IIS restriction endonucleases, such as *BsaI* and *BsmBI*, which cleave outside their recognition sequence, generating a user-defined four-base overhang and leaving no scar after ligation **(Figure 1A)**.

Our method assembles short fragment pools into a longer DNA library. Given the 256 possible four-base overhangs, ideal Watson-Crick base pairing allows for a maximum of 128 pairs of short fragments to be ligated simultaneously via unique complementary overhangs. However, considering the potential of mismatches between overhangs and the convenience of downstream PCR retrieval in 96-well plates, we adjusted our expectations. Ideally, we aim to assemble 96 DNA sequences in a single-pot reaction **(Figure 1B)**. Additionally, this process is fully modular and repeatable, achieved by introducing multiple Type IIS restriction enzymes or blocking digestion through methylation. This enables the hierarchical assembly of final DNA sequences from 4, 8, or even more fragments, resulting in exceptionally long sequences. Moreover, all sequences could be retrieved conveniently by PCR using 96 pairs of invariable sequence barcode primers **(Figure 1C)**.

In practice, we designed the oligo-pools as illustrated in **Figure 1C** to evaluate the performance of our method. These oligo-pools were subsequently synthesized by Twist Bioscience.

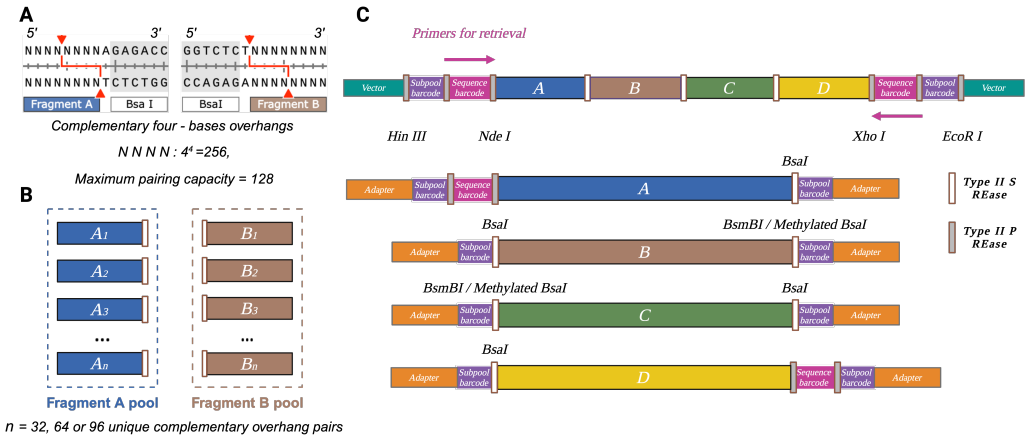


Figure 1: Schematic illustration of the assembly process and design of the oligo-pools. A, Complementary overhangs exposed after *BsaI* digestion. B, Assembly of two short fragment pools into a longer DNA library. Ideally, our aim is to assemble 96 pairs of short fragments simultaneously through unique complementary overhangs. C, Design strategy of the oligo-pools: Adapter, for amplifying the oligo-pools; Subpool barcode, for extracting individual fragment subpools from the oligo-pools; Sequence barcode, for retrieving each sequence from the assembled library.

3 RESULTS AND DISCUSSION

3.1 HIGH-FIDELITY ASSEMBLY OF A 32-GENE LIBRARY (˜600 BP EACH)

To validate the practicability of this method, we began by assembling a library of moderate complexity, comprising 32 genes, each approximately 600 bp in length. These genes were bioinformatically partitioned into four fragments, with careful attention to ensure that the usage of complementary overhang pairs was unique at each assembly step, thereby eliminating the mispairing between fragments.

Figure 2A outlines the steps of the assembly process. Initially, four fragment subpools were extracted from the oligo-pools using respective subpool barcode primers. Subsequently, the purified A and B subpools, along with the C and D subpools, underwent *BsaI* digestion and ligation with T4 ligase. It is evident that two short fragment subpools were assembled into a longer one. Next, we amplified the longer fragments AB and CD through PCR for the next round of assembly. Meanwhile, we also included two negative controls by amplifying the unassembled fragments under the same conditions. In contrast, no bands appeared in the same location, suggesting that the bands of AB and CD were indeed the successfully assembled products. After that, the purified AB and CD fragments were subjected to *BsmBI* digestion and ligation again. Finally, the assembled DNA library (ABCD) was amplified using the same principle.

To assess the quality of our assembly, we analyzed the assembled DNA library using PacBio HiFi sequencing **(Figure 2B)**. An unprecedented 58.78% of the assembly products were error-free, with all 32 target sequences represented. Additionally, 30.78% of the products carried small mismatches or indels, which are likely introduced during oligo synthesis or PCR by polymerase. Therefore, a total of 89.56% of the products were correctly assembled. Only 10.45% of the products were chimeras resulting from mismatch between overhangs.

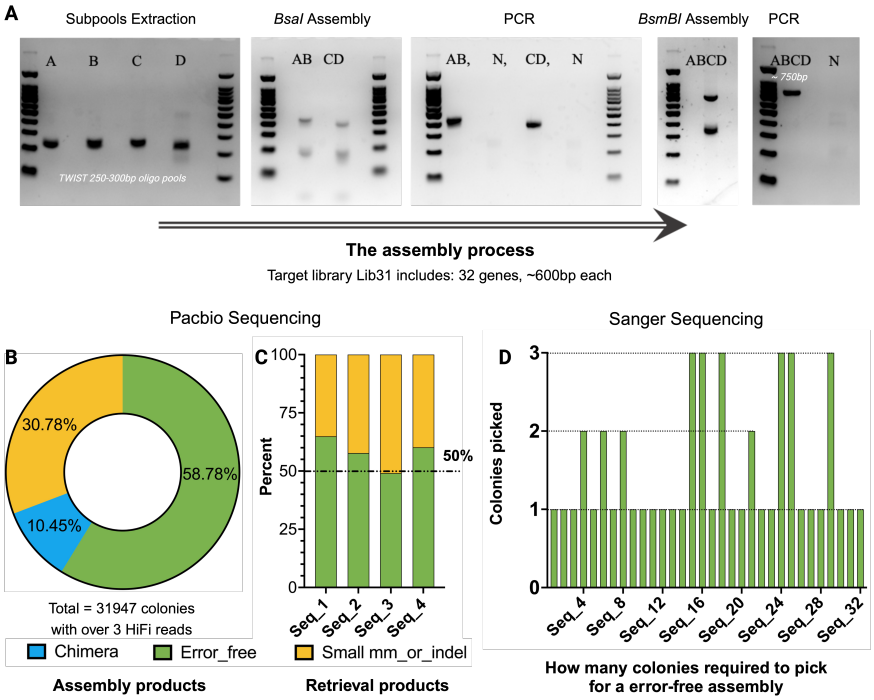


Figure 2: The assembly process and sequencing results of a target library Lib31 (32 genes, ˜600bp each). A, Gel image of the assembly process. B, Pacbio sequencing results of the assembly product. C, Pacbio sequencing results of four retrieved genes. D, Number of colonies required to pick for Sanger sequencing to obtain an error-free DNA sequence.

It is worth noting that these were the products after three assembly processes, so we can calculate the rate of correct assembly—denoted as 'p'—for a single assembly using the equation below:

$$p^3 = 0.8956$$

Thus, approximately 96.4% of the products after a single assembly were correctly ligated, which is remarkably high and indicates the exceptional fidelity of our method. Meanwhile, the chimera rate of a single assembly is only 3.6%.

Furthermore, we successfully retrieved all 32 target genes from the library using sequence barcode primers through PCR. PacBio sequencing results revealed that more than half of the retrieval products for the first four genes were error-free **(Figure 2C)**. Additionally, fewer than three colonies were required for Sanger sequencing to obtain an error-free assembly of each target gene **(Figure 2D)**. Overall, Only 48 colonies were sequenced to obtain error-free sequences for all 32 target genes.

3.2 Error-Correction Enzymes Enable the Assembly of Up to 96 Genes

[1]

3.2.1 Assembly Fidelity Decreases with Increasing Complexity

Next, we investigated the impact of DNA library complexity on assembly fidelity. 96 genes, each 600 bp in length, were partitioned into four fragments concurrently. These genes were then split into three DNA libraries (Lib38, 39 and 40), each containing 32 genes. This allowed us to increase the complexity by adding one additional library at a time.

We assembled these libraries with increasing complexity following the same protocol outlined earlier. **Table 1** shows the Sanger sequencing results of the assembly with increasing complexity. The ratio of chimera increased slightly from 8.3% to 16.7% when assembling 64 genes, and then surged to 43.8% when assembling 96 genes, indicating a higher incidence of mismatches between overhangs. In this case, error-free assembly products became increasingly rare.

Table 1: Sanger sequencing results for the assembly with rising complexity

	Lib38	Lib38 and 39	Lib38, 39 and 40
Complexity	32 genes	64 genes	96 genes
Error-free	13/24 (54.2%)	11/24 (45.8%)	6/32 (18.8%)
Small_mm_or_indel	9/24 (37.5%)	9/24 (37.5%)	12/32 (37.5%)
Chimera	**2/24 (8.3%)**	**4/24 (16.7%)**	**14/32 (43.8%)**

3.2.2 Error-correction enzymes enhance fidelity and diminish point mutations introduced during oligo synthesis

Given that the issue arises from the higher incidence of mismatches between overhangs, we aim to address this by cleaving them with error-correction enzymes, such as Authenticase (NEB).

Authenticase is a mixture of nucleases that recognize and cleave mismatches and indels **(Figure 3A)**. If two short fragments are incorrectly ligated due to overhang mismatch, Authenticase will cleave the chimeric product at the mismatch, preventing its amplification in the next round and thereby eliminating chimeras. Additionally, Authenticase may help diminish point mutations introduced during oligo synthesis. Because there are more error-free fragments in the extracted subpools, with an additional round of denaturation and reannealing, fragments with mutations are more likely to pair with error-free ones, forming heteroduplex DNA with mismatches or indels. Authenticase will then cleave these heteroduplexes as well, further diminishing point mutations.

Table 2 shows the Sanger sequencing results for the assembly of another DNA library (96 genes, ~600bp), either with or without Authenticase digestion. The results suggest that the ratio of chimera in the assembly product dramatically decreased from 37.5% to 13.6%, while the error-free ratio

[1] Pacbio sequencing results are not ready yet for this section, but they will be prepared at the moment of the workshop.

increased from 41.6% to 63.6%. Furthermore, **Figure 3B** indicates that the majority of the ligation products were digested by Authenticase, which would not be the case if only the chimeric products were digested, strongly supporting our hypothesis. All the genes within this library have also been retrieved conveniently.**(Supplemental_Figure_1)**

In this case, Authenticase maintaioned the similar proportion of error-free products while assembling 96 genes.

Table 2: Sanger sequencing results for the assembly, either with or without Authenticase digestion

	Lib08 without Authenticase digestion	Lib08 with Authenticase digestion
Complexity	96 genes	96 genes
Error-free	10/24 (41.6%)	14/22 (63.6%)
Small_mm_or_indel	5/24 (20.8%)	5/22 (22.7%)
Chimera	**9/24 (37.5%)**	**3/22 (13.6%)**

3.3 Blocking *BsaI* digestion through methylation may enable modular and hierarchical assembly

We have achieved the two-step assembly of four fragments by alternatively using *BsaI* and *BsmBI*. However, multi-step hierarchical assembly, for example, a three-step assembly of eight fragments, would necessitate various Type IIS restriction enzymes, which would impose additional constraints on the DNA fragments to be assembled, requiring the removal of multiple internal restriction sites.

To address this issue, we hypothesized that by replacing *BsmBI* recognition sites with methylated *BsaI* recognition sites, the corresponding ends would be protected from digestion, preventing these overhangs from interfering with the assembly process.

To validate this hypothesis, we synthesized a B fragment with a structure similar to the oligo-pools, the only difference being the replacement of *BsmBI* with *BsaI* **(Figure 3C)**. We then amplified this fragment using the 3' primer, either containing a methylated *BsaI* recognition site or not. The amplified product was subsequently subjected to overnight *BsaI* digestion. Afterward, we analyzed the digestion products by capillary electrophoresis. The results clearly indicated that only the end without methylation was digested, while the methylated end remained completely protected **(Figure 3D)**.

In this case, both the applicability and final length of the DNA library to be assembled using our method are significantly extended.

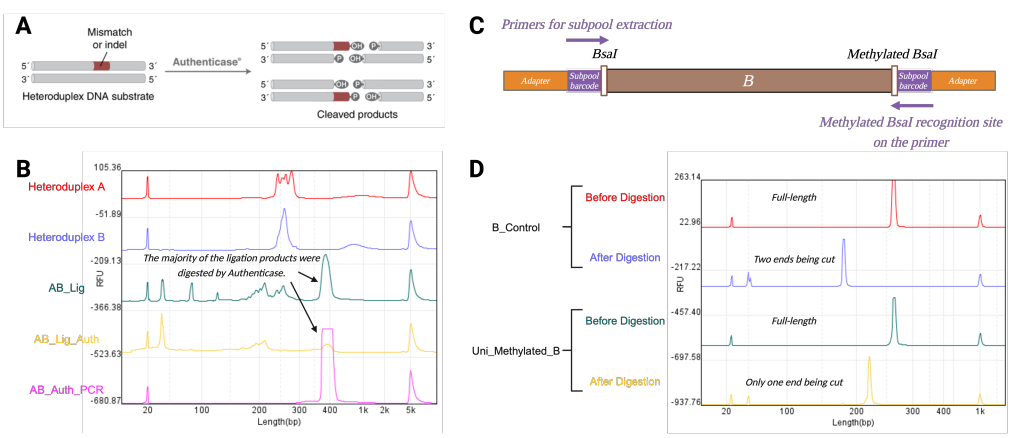


Figure 3: Error correction of the assembly products. A, Authenticase (NEB) cleaves double-stranded DNA (dsDNA) carrying mismatches or indels. B, Capillary electrophoresis analysis of the heteroduplex A and B fragments. AB ligation product (AB_Lig), AB ligation product after Authenticase digestion (AB_Lig_Auth) and PCR-amplified AB fragment (AB_Auth_PCR). C, Diagram of blocking *BsaI* digestion through methylation. D, Capillary electrophoresis analysis of the B_Control and Uni_Methylated_B fragments before and after *BsaI* digestion.

REFERENCES

Alex Hoose, Richard Vellacott, Marko Storch, Paul S. Freemont, and Maxim G. Ryadnov. Dna synthesis technologies to close the gene writing gap. *Nature Reviews Chemistry*, 7(3):144–161, January 2023. doi: 10.1038/s41570-022-00456-9. URL https://doi.org/10.1038/s41570-022-00456-9.

Jason C. Klein, Marc J. Lajoie, Jerrod J. Schwartz, Eva-Maria Strauch, Jorgen Nelson, David Baker, and Jay Shendure. Multiplex pairwise assembly of array-derived dna oligonucleotides. *Nucleic Acids Research*, 44(5):e43–e43, March 2016. doi: 10.1093/nar/gkv1177. URL https://doi.org/10.1093/nar/gkv1177.

Sriram Kosuri, Nikolai Eroshenko, Emily M LeProust, Michael Super, Jeffrey Way, Jin Billy Li, and George M. Church. Scalable gene synthesis by selective amplification of dna pools from high-fidelity microchips. *Nature Biotechnology*, 28(12):1295–1299, December 2010. doi: 10.1038/nbt.1716. URL https://doi.org/10.1038/nbt.1716.

Ziva Moravec, Yue Zhao, Rhianne Voogd, Danielle R. Cook, Seon Kinrot, Benjamin Capra, Haiyan Yang, and et al. Discovery of tumor-reactive t cell receptors by massively parallel library synthesis and screening. *Nature Biotechnology*, April 2024. doi: 10.1038/s41587-024-02210-6. URL https://doi.org/10.1038/s41587-024-02210-6.

Calin Plesa, Angus M. Sidore, Nathan B. Lubock, Di Zhang, and Sriram Kosuri. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science*, 359(6373):343–347, January 2018. doi: 10.1126/science.aao5167. URL https://doi.org/10.1126/science.aao5167.

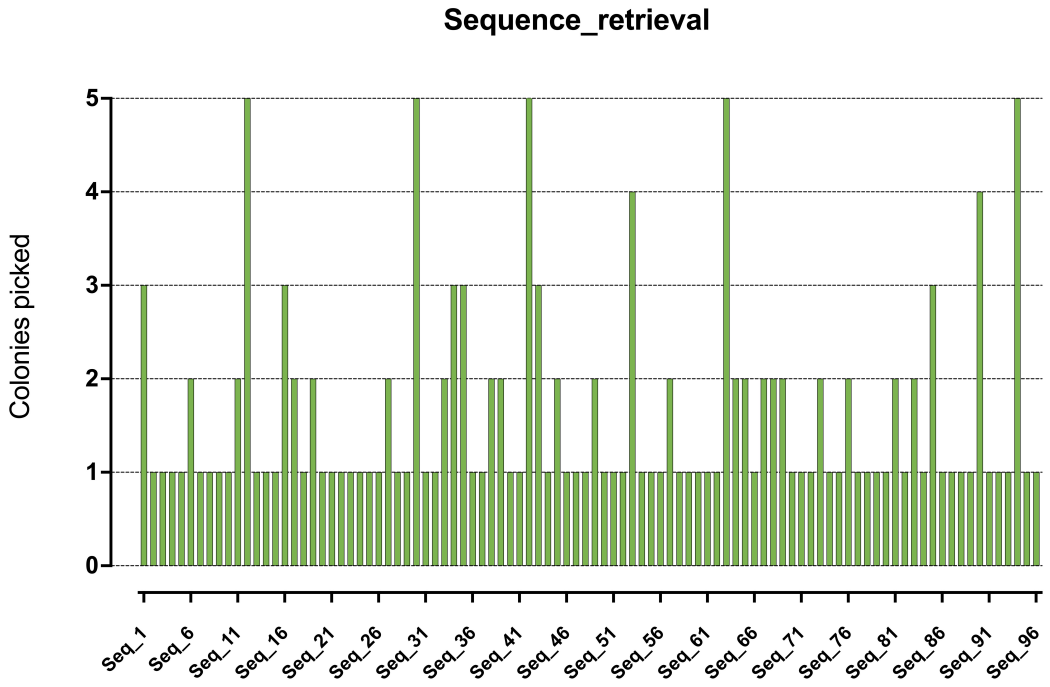
Jiayuan Quan, Ishtiaq Saaem, Nicholas Tang, Siying Ma, Nicolas Negre, Hui Gong, Kevin P. White, and Jingdong Tian. Parallel on-chip gene synthesis and application to optimization of protein expression. *Nature Biotechnology*, 29(5):449–452, May 2011. doi: 10.1038/nbt.1847. URL https://doi.org/10.1038/nbt.1847.

Ishtiaq Saaem, Siying Ma, Jiayuan Quan, and Jingdong Tian. Error correction of microchip synthesized genes using surveyor nuclease. *Nucleic Acids Research*, 40(3):e23–e23, February 2012. doi: 10.1093/nar/gkr887. URL https://doi.org/10.1093/nar/gkr887.

Ana Filipa Sequeira, Catarina I. P. D. Guerreiro, Renaud Vincentelli, and Carlos M. G. A. Fontes. T7 endonuclease i mediates error correction in artificial gene synthesis. *Molecular Biotechnology*, 58 (8):573–584, September 2016. doi: 10.1007/s12033-016-9957-7. URL https://doi.org/10.1007/s12033-016-9957-7.

Jingdong Tian, Hui Gong, Nijing Sheng, Xiaochuan Zhou, Erdogan Gulari, Xiaolian Gao, and George Church. Accurate multiplex gene synthesis from programmable dna microchips. *Nature*, 432(7020):1050–1054, December 2004. doi: 10.1038/nature03151. URL https://doi.org/10.1038/nature03151.

A SUPPLEMENTAL FIGURE



Supplemental_Figure_1: Overall 155 colonies were picked for Sanger sequencing to retrieve all 96 error-free DNA sequences from DNA library Lib08, which has been Authenticase digested to improve fidelity and exclude point mutations