HIERARCHICAL MULTIPLEX PAIRWISE GOLDEN GATE ASSEMBLY: CONVERTING SHORT OLIGO-POOLS INTO LONGER DNA LIBRARIES

Shaozhong Zou[1], Zhien Wu[1,2] & Chunfu Xu[1,3]
[1]National Institute of Biological Science, Beijing, China
[2]Peking University–Tsinghua University–National Institute of Biological Sciences (PTN)
Joint Graduate Program, School of Life Sciences, Tsinghua University, Beijing, China
[3]Tsinghua Institute of Multidisciplinary Biomedical Research, Tsinghua University, Beijing, China
{zoushaozhong,wuzhien,xuchunfu}@nibs.ac.cn

ABSTRACT

Large-scale screening and high-throughput experimental data generation are essential for advancing AI-driven biomolecular design. However, these processes are economically unfeasible due to the high costs associated with synthesizing gene-sized DNA sequences at scale. 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. In contrast to 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. If numerous reactions are conducted in parallel—for example, in a 96-well plate—we can readily assemble up to 9,216 (96×96) genes. When combined with advances in automation technologies, this enables the efficient and cost-effective synthesis of gene-sized DNA sequences at scale, thereby accelerating the generation of experimental data for the biomolecular design community.

1 INTRODUCTION

In recent years, DNA synthesis technologies have played a pivotal role in advancing numerous research areas, including AI-driven biomolecular design. For a given design task, typically only a small number of candidate sequences with top benchmarking results are synthesized for experimental validation, significantly limiting the sequence space being explored. In this context, large-scale screening and high-throughput experimental data generation are essential for advancing AI-driven biomolecular design. However, their high cost associated with the large-scale synthesis of gene-sized DNA sequences make these process economically unfeasible.

To synthesize a gene-sized DNA sequence, it is first split into short oligonucleotides with complementary sequences. 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; Lund et al., 2024) ; another approach is to assemble DNA sequences in a pool 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 ($> 50\%$). Furthermore, an additional high-throughput sequencing step is generally required to retrieve an error-free sequence.

To address this challenge, we developed a novel method for assembling short oligo-pools into longer DNA libraries. In contrast to Polymerase Cycling Assembly (PCA), we employed Golden Gate Assembly (GGA) to precisely ligate short DNA fragments. Using this approach, We successfully assembled a set of high-quality DNA libraries, containing up to 96 gene-sized sequences (600 bp), with more than 50% of products being error-free. Additionally, we have efficiently retrieved all 96 error-free DNA sequences from the resulting DNA library by picking only 155 colonies for Sanger sequencing. If numerous reactions are conducted in parallel—for example, in a 96-well plate—we can readily assemble up to 9,216 (96×96) genes. When combined with advances in automation technologies, this enables the efficient and cost-effective synthesis of gene-sized DNA sequences at scale, thereby accelerating the generation of experimental data for the biomolecular design community.

2 METHODS:

Golden Gate Assembly (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)**. Traditionally, Golden Gate Assembly (GGA) has been used to assemble multiple DNA fragments into a single linear construct, with the order predetermined by complementary overhang pairs (Pryor et al., 2022). However, we shifted this paradigm by enabling multiple pairwise ligation events in a single-pot reaction. In this context, if the pairing between complementary overhangs is both unique and orthogonal(Potapov et al., 2018), then parallel pairwise ligation of fragments from multiple genes can be achieved with equally high accuracy **(Figure 1B)**.

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 and the convenience of downstream PCR retrieval in 96-well plates, we adjusted our expectations. We expect to assemble up to 96 DNA sequences in a single-pot reaction. Meanwhile, by introducing multiple Type IIS restriction enzymes or block-

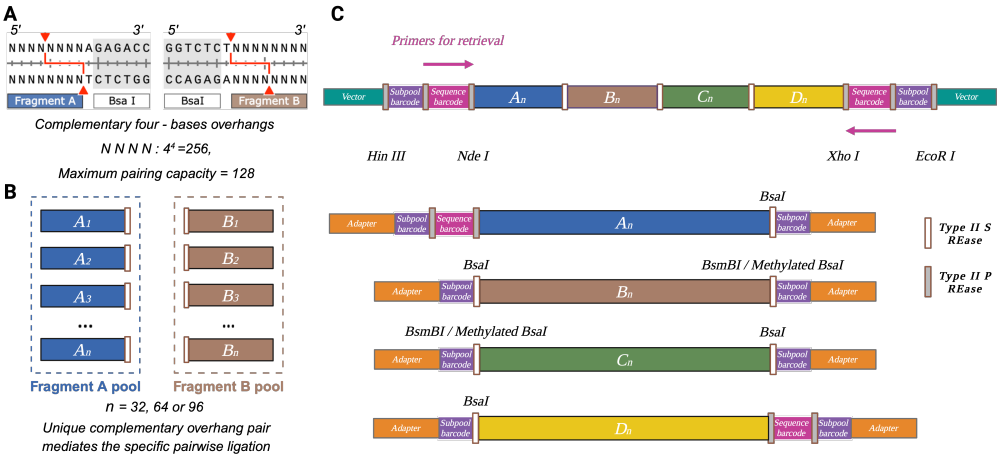


Figure 1: Schematic illustration of the assembly process and design of the oligo-pools. A, Complementary overhangs exposed after *BsaI* digestion. B, Assembly of two short fragment pools into a longer DNA library. We expect to assemble up to 96 pairs of short fragments simultaneously through unique complementary overhangs pairs. 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.

ing digestion through methylation, hierarchical assembly from 4, 8, or even more fragments can be achieved, producing exceptionally long sequences. Moreover, all sequences can be retrieved conveniently by PCR using 96 pairs of fixed 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 purcahsed from Twist Bioscience.

3 RESULTS

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 primers binding to respective subpool barcodes. Subsequently, the purified subpools A and B, along with subpools C and D, underwent *BsaI* digestion and ligation with T4 ligase. It is evident that two short fragment pools 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.

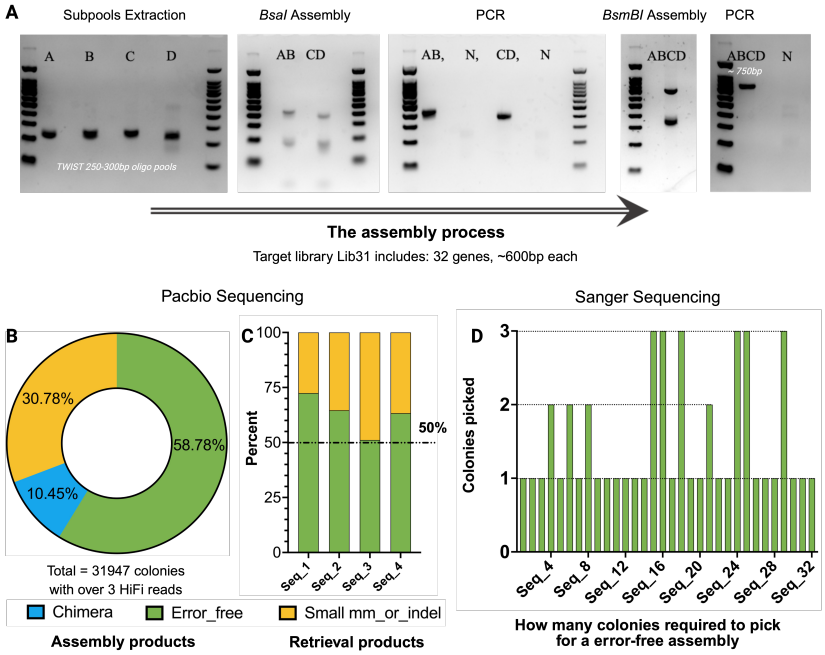


Figure 2: The assembly process and sequencing results of a target library Lib31 (32 genes, ˜600bp each). A, Gel image of the assembly process. B, Pacbio sequencing results of the assembly product. C, Pacbio sequencing results of four retrieved genes. Number of colonies analyzed: Seq_1 = 1683, Seq_2 = 277, Seq_3 = 401, Seq_4 = 338. D, Number of colonies picked for Sanger sequencing to obtain an error-free DNA sequence.

To assess the quality of our assembly, the assembled DNA library was cloned into the pUC19 vector, transformed into E. coli Top10 cells, and the resulting colonies were analyzed using PacBio HiFi sequencing. **(Supplemental_Figure_1)**

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-pools 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 2B)**.

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

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 to pick 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

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-pools synthesis. Because there are more error-free fragments in the extracted subpools, with an additional round of denaturation and reannealing, single strand with mutations are more likely to pair with error-free ones, forming heteroduplex 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 significantly decreased from 37.5% to 13.6%, while the error-free ratio 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_2)**

In this case, Authenticase maintained the similarly high proportion of error-free products, even 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 replacing the *BsmBI* recognition sites with methylated *BsaI* recognition sites would protect the corresponding ends from digestion, thereby preventing these overhangs from interfering with the ongoing assembly process at the other end.

To validate this hypothesis, we synthesized a DNA with similar structure to the B fragment, the only difference being the replacement of *BsmBI* with *BsaI* **(Figure 3C)**. We then amplified this fragment using the 3' primers, 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)**.

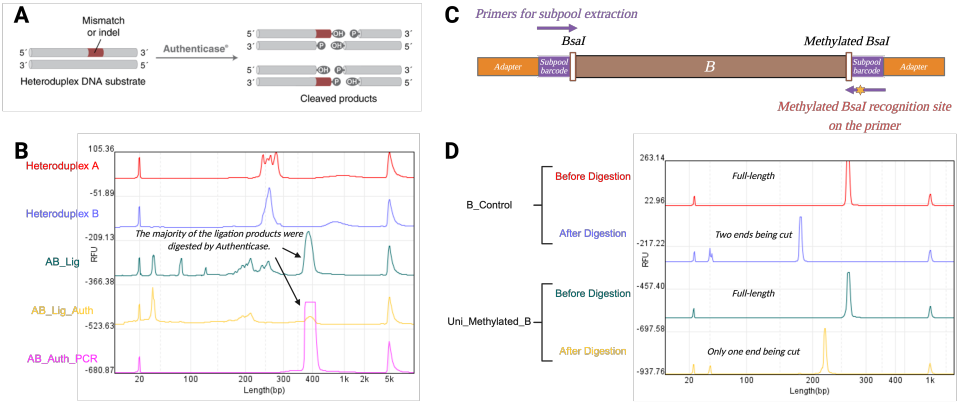


Figure 3: Error correction of the assembly products. A, Authenticase (NEB) cleaves double-stranded DNA (dsDNA) carrying mismatches or indels. B, Capillary electrophoresis analysis of the heteroduplex A and B fragments. AB ligation product (AB_Lig), AB ligation product after Authenticase digestion (AB_Lig_Auth) 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.

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

4 DISSCUSSION:

The massive scalability and minimal per-sequence reagent consumption of chip-synthesized oligo-pools have long made them a highly attractive route for reducing the cost of DNA synthesis. However, efficiently converting them into error-free, gene-sized DNA sequences remains challenging due to their short length and high sequence complexity.

Selective separation of the oligo-pool into subpools is a cornerstone of the strategy(Kosuri et al., 2010; Quan et al., 2011), as it reduces complexity and thereby enhances assembly success through Polymerase Cycling Assembly (PCA). Building on this strategy, recent work (Lund et al., 2024) has suggested that Golden Gate Assembly (GGA) can also be employed to assemble these separated oligos as an alternative to the PCA approach. However, all of these studies were limited to assembling only a single gene in each single-pot reaction, which incurred considerable assembly costs and hindered scalability.

In this context, we shifted the paradigm of Golden Gate Assembly (GGA) from assembling a single linear construct to enabling multiple pairwise ligation events within a single-pot reaction. By introducing multiple Type IIS restriction enzymes or blocking digestion through methylation, we can assemble exceptionally long DNA sequences from 4, 8, or even more fragments.

We restricted it to linearly pairwise ligation, primarily due to concerns about ligation efficiency. For example, when assembling n fragments linearly with a pairwise ligation efficiency of e, the overall ligation efficiency becomes e^n, which decays exponentially as the number of fragments increases. This significantly limits the number of correctly assembled molecules. Therefore, after each step of pairwise ligation, we re-amplified the ligation products to compensate for the loss in ligation efficiency.

In this work, starting from 300 bp oligo-pools purchased from TWIST Bioscience, we successfully assembled a set of high-quality DNA libraries containing up to 96 gene-sized sequences (600 bp) through a two-step hierarchical assembly of four fragments, with more than 50% of the products being error-free. Moreover, although not yet experimentally verified, the use of longer oligos could enable the assembly of even longer DNA libraries.

Additionally, all sequences can be conveniently retrieved by PCR using 96 pairs of fixed sequence barcode primers, making the approach broadly applicable to DNA synthesis. Once a DNA library is assembled, individual sequences can be efficiently retrieved by PCR in a 96-well plate, with specific primer sets preloaded in each well.

By assembling DNA sequences in a pooled format, we have significantly reduced the average assembly cost and thereby improved scalability. If numerous assembly reactions are conducted in parallel—for example, in a 96-well plate—we can readily assemble up to 9,216 (96×96) genes. When combined with advances in automation technologies, this enables the efficient and cost-effective synthesis of gene-sized DNA sequences at scale, thereby accelerating the generation of experimental data for the biomolecular design community.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Science and Technology of the People's Republic of China (Project 2022YFA1303703). We thank Julian Jude of TWIST Bioscience for the valuable suggestion regarding the use of methylation to block digestion. Shaozhong Zou would like to thank Prof. Jia Zheng (Westlake University) and Prof. Zhen Liu (National Institute of Biological Sciences, Beijing) for the training he received in their laboratories.

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.

Sean Lund, Vladimir Potapov, Sean R. Johnson, Jackson Buss, and Nathan A. Tanner. Highly parallelized construction of dna from low-cost oligonucleotide mixtures using data-optimized assembly design and golden gate. *ACS Synthetic Biology*, 13(3):745–751, March 2024. doi: 10.1021/acssynbio.3c00694. URL https://doi.org/10.1021/acssynbio.3c00694.

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.

Vladimir Potapov, Jennifer L. Ong, Rebecca B. Kucera, Bradley W. Langhorst, Katharina Bilotti, John M. Pryor, Eric J. Cantor, et al. Comprehensive profiling of four base overhang ligation fidelity by t4 dna ligase and application to dna assembly. *ACS Synthetic Biology*, 7(11):2665–2674, November 2018. doi: 10.1021/acssynbio.8b00333. URL https://doi.org/10.1021/acssynbio.8b00333.

John M. Pryor, Vladimir Potapov, Katharina Bilotti, Nilisha Pokhrel, and Gregory J. S. Lohman. Rapid 40 kb genome construction from 52 parts through data-optimized assembly design. *ACS Synthetic Biology*, 11(6):2036–2042, June 2022. doi: 10.1021/acssynbio.1c00525. URL https://doi.org/10.1021/acssynbio.1c00525.

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.

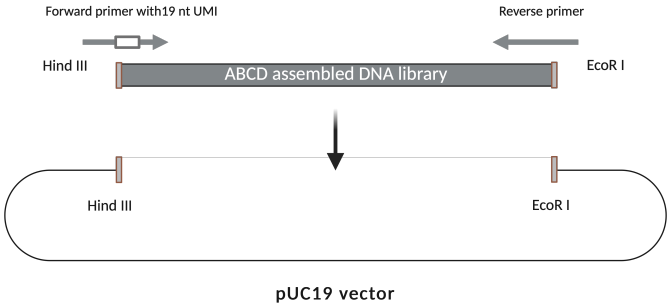
Jingdong Tian, Hui Gong, Nijing Sheng, Xiaochuan Zhou, Erdogan Gulari, Xiaolian Gao, and George Church. Accurate multiplex gene synthesis from programmable dna microchips. *Nature*, 432(7020):1050–1054, December 2004. doi: 10.1038/nature03151. URL https://doi.org/10.1038/nature03151.

A SUPPLEMENTAL MATERIAL

After amplification and purification, the assembled DNA libraries (ABCD) were used in three ways: (1) digested with HindIII and EcoRI and ligated into the pUC19 vector for PacBio sequencing analysis, (2) used as a template for retrieving individual sequences from the library, and (3) digested with NdeI and XhoI for experimental functional screening.

B STRATEGY FOR PACBIO SEQUENCING

The assembled DNA library was PCR-amplified using primers containing a 19-nt (15-nt for individual sequence retrieval) fully degenerate unique molecular identifiers (UMI) to uniquely label each template molecule. Amplicons were digested with HindIII and EcoRI (via primer-introduced sites) and ligated into the pUC19 vector. The ligation products were transformed into in-house–prepared chemically competent E. coli TOP10 cells and selected on carbenicillin LB plates, with each colony carrying a uniquely UMI-tagged insert. Colonies were pooled for plasmid extraction, and the target region was re-amplified for PacBio HiFi circular consensus sequencing. Colony numbers were controlled to ensure each UMI was represented by ≥ 3 reads. During data processing, only UMI groups supported by ≥ 3 reads were used for consensus generation, filtering out low-confidence sequences and ensuring high-accuracy reconstruction of original molecules.

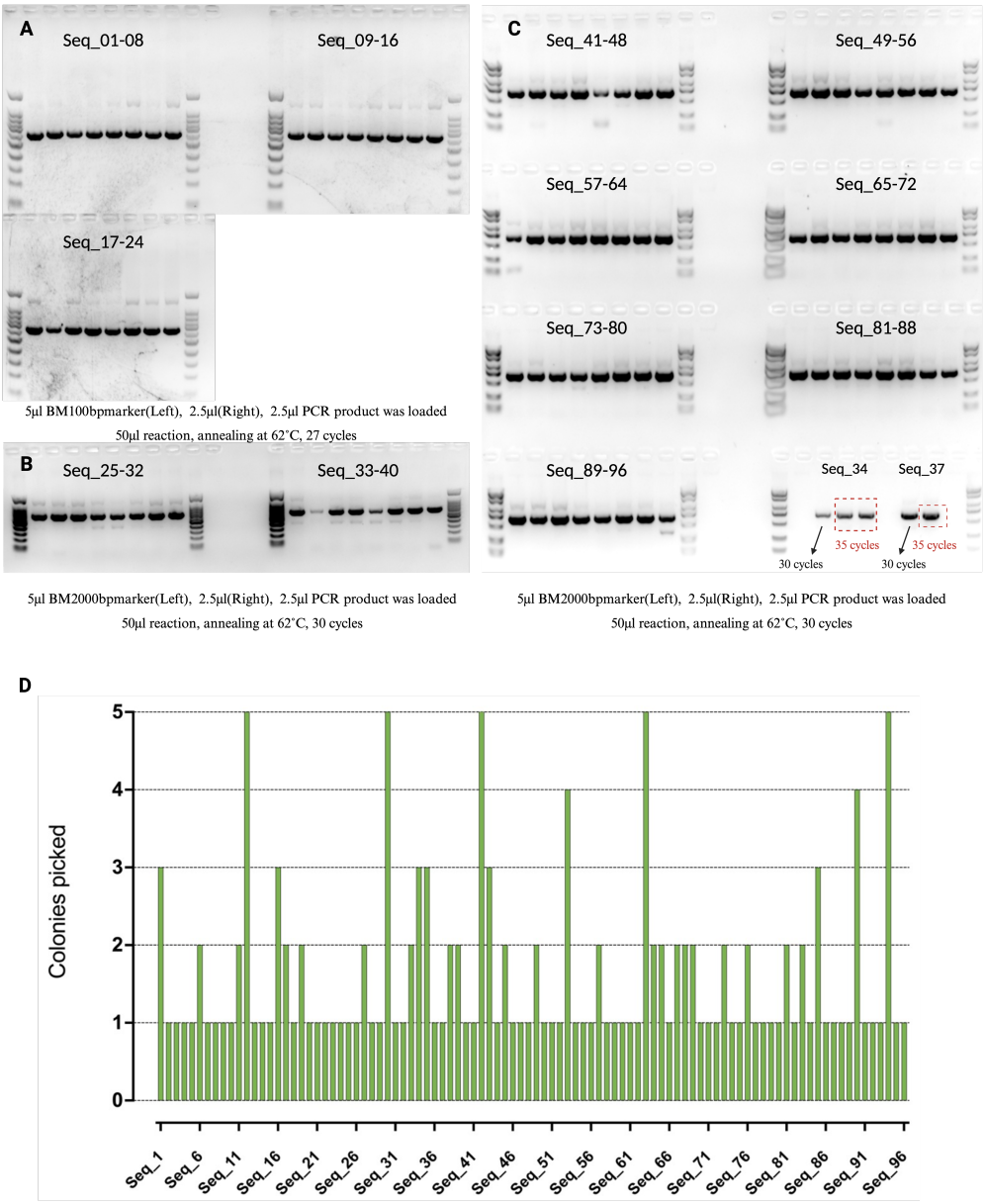


Supplemental_Figure_1: Strategy for PacBio sequencing. A 19-nt UMI uniquely labels each template molecule, which is then ligated into the pUC19 vector. Clonal copies were generated using the bacterial DNA replication machinery, and subsequently subjected to PacBio HiFi sequencing.

C INDIVIDUAL SEQUENCE RETRIEVAL FROM RESULTING DNA LIBRARY

Using equal amounts of the assembled DNA library (Lib08 treated with Authenticase digestion) as the template, we successfully retrieved all 96 genes via PCR, employing 96 primer pairs targeting their respective sequence barcodes. The results of these PCR reactions are shown in the Supplemental Figure below. Most amplicons were generated in a single PCR round. An additional round of PCR with five extra cycles was performed only for Seq_34 and Seq_37 due to concerns about their initially weak bands, which suggested low yield. In general, all PCR products displayed sharp and clean bands.

All 96 PCR reactions were purified using magnetic beads. Following digestion with HindIII and EcoRI, the products were ligated into the pUC19 vector, transformed into E. coli TOP10 cells and selected on carbenicillin LB plates as well. Several colonies were picked for Sanger sequencing as shown in Supplemental Figure D, while others were subjected to PacBio sequencing as described above. In total, 155 colonies were picked for Sanger sequencing to retrieve all 96 error-free DNA sequences. Notably, approximately 15% of the colonies carried empty vectors lacking inserts, likely due to incomplete digestion of the pUC19 backbone. Therefore, in principle, fewer than 155 colonies would have been sufficient to retrieve all 96 error-free sequences.



Supplemental_Figure_2: Individual sequence retrieval from resulting DNA library. A, B, C, Gel image of retrieval PCR products. D, Number of colonies pciked for Sanger sequencing to retrieve all 96 error-free DNA sequences from the assembled DNA library (Lib08 treated with Authenticase digestion).