

Massively parallel homogeneous amplification of chip-scale DNA for DNA information storage (MPHAC-DIS)

Received: 7 August 2024

Accepted: 2 January 2025

Published online: 14 January 2025

Zhi Weng^{1,4}, Jiangxue Li^{1,4}, Yi Wu¹, Xuehao Xiu¹, Fei Wang², Xiaolei Zuo^{2,3}✉, Ping Song¹✉ & Chunhai Fan²✉

Chip scale DNA synthesis offers a high-throughput and cost-effective method for large-scale DNA-based information storage. Nevertheless, unbiased information retrieval from low-copy-number sequences remains a barricade that largely arises from the indispensable DNA amplification. Here, we devise a simulation-guided quantitative primer-template hybridization strategy to realize massively parallel homogeneous amplification of chip-scale DNA for DNA information storage (MPHAC-DIS). Using a fixed-energy primer design, we demonstrate the unbiasedness of MPHAC for amplifying 100,000-plex sequences. Simulations reveal that MPHAC achieves a fold-80 value of 1.0 compared to 3.2 with conventional fixed-length primers, lowering costs by up to four orders of magnitude through reduced over-sequencing. The MPHAC-DIS system using 35,406 encoded oligonucleotide allows simultaneous access of multimedia files including text, images, and videos with high decoding accuracy at very low sequencing depths. Specifically, even a $\sim 1\times$ sequencing depth, with the combination of machine learning, results in an acceptable decoding accuracy of $\sim 80\%$. The programmable and predictable MPHAC-DIS method thus opens new door for DNA-based large-scale data storage with potential industrial applications.

DNA-based information storage has emerged as a compelling alternative for high-capacity data storage due to its exceptional encoding capabilities, long-term storage potential, and an impressive storage density surpassing 17 EB/g^{1–4}. This method involves encoding data into DNA sequences through DNA synthesis (writing) and retrieving the data via DNA sequencing (reading). Therein, DNA synthesis includes chemical column-based and chip-based method. Chemical column-based DNA synthesis produces high quantities per oligonucleotide (in picomoles) but offers limited diversity throughput, typically ranging from 96 to 1536 oligonucleotides, which poses a significant bottleneck for high-throughput sequence synthesis for practical DNA storage

applications. For instance, storing 1 terabyte of data would require $\sim 10^{10}$ oligonucleotides with 150 nucleotides (nt) long.

In contrast, chip-scale DNA synthesis, a burgeoning industry, achieves the simultaneous synthesis of 25 million molecules/cm², amounting to $\sim 8.4 \times 10^6$ oligonucleotides^{5,6}. This method also benefits from a substantial cost reduction, ~ 1000 -fold (\$0.0001/base vs. \$0.1/base), making it favorable for DNA information storage^{7,8}. However, the number of copies of a single sequence synthesized by this method typically ranges from 10^5 to 10^{12} copies, with concentrations in the femtomolar range, insufficient for commercial second-generation sequencing platforms (e.g., Illumina) which require nanomolar levels

¹School of Biomedical Engineering, Zhangjiang Institute for Advanced Study and National Center for Translational Medicine, Shanghai Jiao Tong University, Shanghai, China. ²School of Chemistry and Chemical Engineering, New Cornerstone Science Laboratory, Frontiers Science Center for Transformative Molecules, National Center for Translational Medicine, Shanghai Jiao Tong University, Shanghai, China. ³Institute of Molecular Medicine, Shanghai Key Laboratory for Nucleic Acid Chemistry and Nanomedicine, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China. ⁴These authors contributed equally: Zhi Weng, Jiangxue Li. ✉ e-mail: zuoxiaolei@sjtu.edu.cn; songpingsjtu@sjtu.edu.cn; fanchunhai@sjtu.edu.cn

of incorporated DNA for information readout^{9–11}. Therefore, achieving bias-free DNA amplification is a critical concern for high-throughput DNA information storage^{12–14}.

The polymerase chain reaction (PCR) is a pivotal technology in molecular biology that can amplify DNA quantities by a billionfold in a short time. This amplification capability has made PCR essential in DNA information storage for amplification and random access^{15–17}. Traditionally, primers are designed at a fixed length, typically 20 nt, to simplify data processing¹⁶. However, this one-size-fits-all approach can lead to variable binding efficiencies, which are exponentially exacerbated across PCR cycles, ultimately compromising the uniformity of the DNA library^{18,19}. The consequences of uneven amplification are multi-fold: (1) the potential loss of a subset of oligonucleotides, rendering them undecodable; (2) increased sequencing costs due to the need for higher coverage of underrepresented oligonucleotides; and (3) distortion of the stored data. Several optimization strategies have been explored to address these issues. One common approach maintains primer length consistency while modulating parameters such as GC content, homopolymer prevalence, and the melting temperature (T_m). However, these constraints may exclude numerous oligonucleotide candidates and may not adequately meet the stringent demands of DNA information storage. An alternative approach involves physically separating multiple single PCR amplifications, such as emulsion PCR, which employs water-in-oil droplets to segregate DNA templates in a single reaction vessel, enabling parallel amplification^{20–22}. However, its intricate workflow and substantial organic solvent requirement hinder large-scale applications.

The binding free energies of programmable probes have been widely studied to achieve binding specificity and sensitivity^{23–27}. Here,

we introduce binding energy controlled massively parallel homogeneous amplification of chip-scale DNA (MPHAC), which ensures highly specific primer hybridization with homogeneous standard Gibbs free energy change (ΔG°), enabling high-throughput, bias-free and selective amplification of target sequences by PCR. To experimentally validate MPHAC, we applied the method to DNA information storage (MPHAC-DIS) for simultaneous access of multimedia files including text, images, and videos with high decoding accuracy at very low sequencing depths.

Results

Overview of MPHAC and MPHAC-DIS

MPHAC technology effectively amplified single-stranded DNA templates synthesized in high-throughput microarrays, to higher concentrations, forming double-stranded DNA templates (Fig. 1a). This amplification was achieved through the careful design and selection of primers to regulate the thermodynamic properties of primer-template hybridization (Fig. 1b). This approach, moving away from fixed primer lengths, enabled homogenized amplification across different primer sets, thus achieving homogeneous amplification levels. Consequently, the method facilitated the high-concentration amplification of synthetic DNA for microarrays.

In DNA information storage, primers were traditionally designed with a fixed length, with 76% adherence to this standard. However, only 19% (4 out of 21) of the designs considered hybridization capability, including factors like T_m and ΔG° (Supplementary Fig. 1 and Supplementary Table 1). These thermodynamic parameters critically influenced the sensitivity and specificity of primer-template hybridization, an aspect previously overlooked in the study of DNA information

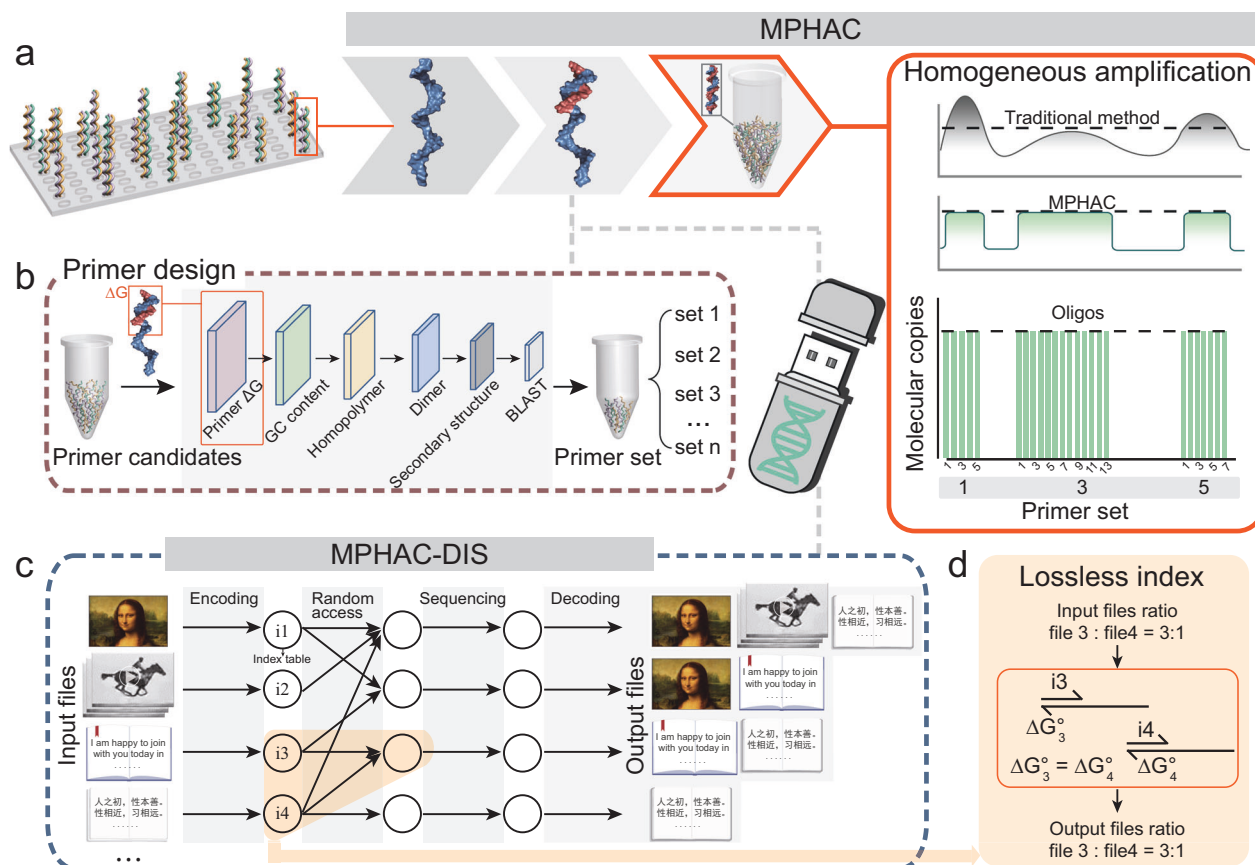


Fig. 1 | Schematic workflow of MPHAC and MPHAC-DIS. a Low concentration single-stranded DNA molecules synthesized on chip were amplified with MPHAC to obtain uniform high-concentration double-stranded DNA molecules. **b** Primer

design and filter workflow. **c** Workflow of multimodal data storage with random access using MPHAC-DIS. **d** Lossless index for DNA information storage to maintain files ratio consistency before and after amplification.

storage. Fixed-length 20 nt primers, without GC content control, exhibited a broad ΔG° range from 0 to $-24 \text{ kcal mol}^{-1}$ and hybridization yields from 0% to 100% (Supplementary Fig. 4a, b). To optimize DNA information storage, a GC content ranging from 45% to 55% was usually utilized^{12,16}, resulting in a mean amplification efficiency of 92.9% for 20 nt primers (Supplementary Fig. 4c). Adjusting the GC content significantly improved amplification efficiency. An extensive screening of 2 million random primers with length ranging from 15 to 30 nt was conducted. Of these, only 8.3% passed the ΔG° filter ($-10.5 \text{ kcal mol}^{-1}$ to $-12.5 \text{ kcal mol}^{-1}$), and a scant 2% conformed to the GC content criteria. Additionally, homopolymers, which could introduce sequencing errors and complicate decoding, necessitated control over their length^{12,28,29}. Although amplification efficiency could be enhanced by extending primer length and controlling GC content (Supplementary Fig. 5 and Supplementary Fig. 6), excessively lengthy primers increased the likelihood of primer dimer formation and the propensity for secondary structures (Supplementary Fig. 7 and Supplementary Fig. 8), interfering with amplification. Hence, controlling the proportion of primer dimers and secondary structures was necessary. Ultimately, after multiple rounds of filtering, a stringent screening of 2 million primer candidates yielded 0.6% that met the criteria. This process produced a library of 11,335 primers, which, as far as we know, was the largest known collection, offering a promising solution for random access in DNA information storage systems (Figs. 1b and Supplementary Fig. 9).

Moreover, the deliberate design of primers was not only used for amplification chip-scale DNA templates but also facilitated random access to files in DNA information storage (Fig. 1c). By integrating different primers with varied data types and creating an index table, various primer sets could be mixed as needed to realize random access to individual or multiple files. These files were then sequenced, decoded, and capable of being saved or restored. Importantly, through precise regulation of primer-template hybridization and amplification capabilities, it was possible to add multiple files proportionally to the index table and accurately read them out (Fig. 1d).

Comparison between MPHAC and conventional fixed length primer amplification

Primer design is a critical aspect of PCR amplification, with primer-template hybridization efficiency serving as a fundamental determinant³⁰. This efficiency can be deduced from the thermodynamic energy ΔG° of hybridization^{25,31,32}. During PCR amplification, any bias in hybridization efficiency is exponentially magnified with each cycle, leading to substantial discrepancies in the number of molecules amplified by different primers. In contrast, consistent hybridization efficiency across primers and templates yields a uniform number of molecules from different templates after identical PCR cycles (Fig. 2a). To illustrate the impact of primer hybridization efficiency on PCR amplification succinctly, we devised a model and simulated the PCR efficiencies for varying hybridization efficiencies of the forward primer (FP) and reverse primer (RP) (Fig. 2b). This model, which considers primer hybridization and extension stages, excludes enzyme activity change on the presumption of its rapid action (Supplementary Fig. 13). To quantify the variation in copy numbers arising from differential amplification efficiencies (E) among primers, we introduced a ratio change ($2^{\text{cycles}}/(1+E)^{\text{cycles}}$). When the template with primers of 100% amplification efficiency is set as the benchmark (value=1), the concentration of a template with primers of 50% amplification efficiency diverges, exemplified by a ratio change of 315 as the cycle number varies (Fig. 2c, d).

To investigate the influence of ΔG° modulation on PCR amplification, we conducted an analysis on two distinct sets of primers, each comprising 100,000 sequences: one set with a fixed-length of 20 nt (FL), and the other with a fixed-energy (FE). The FL-primers controlled GC contents ranging from 40%–60%, while the FE-primers maintained

hybridization thermodynamic energies around $-11.5 \text{ kcal mol}^{-1}$ (Fig. 2e). The ΔG° range for the FL-primers was substantially wider than that of the FE-primers, introducing considerable heterogeneity into the hybridization and amplification processes. Sequencing coverage uniformity was assessed using the fold-80 metric, with values approaching 1 indicating greater homogeneity³³. The coverage comparison from simulated sequencing (Fig. 2f, g) revealed that the FE-primers exhibited better homogeneity (fold-80 = 1), whereas the FL-primers showed poorer uniformity (fold-80 = 3.2). Accurately sequencing low-coverage regions requires additional reads, necessitating higher sequencing depth. Using a standard 30× coverage, this would result in an additional sequencing cost of 11,643-fold (Fig. 2f, g).

Furthermore, employing the 100,000 primers for different combinations, the ratio of corresponding sequences pre- and post-amplification after 20 cycles was calculated. The ratio for FE-primers post-amplification matched the anticipated pattern ($r^2=1.00$), while the ratio for FL-primers deviated substantially ($r^2=0.52$) (Fig. 2h). A systematic comparison of homogeneity, sequence-to-sequence amplification variation (ratio), and sequencing cost for FL- and FE-primers demonstrated that fixed-energy primers offered superior sequence fidelity and reduced costs due to homogeneous amplification (Fig. 2i).

Parameters optimization of MPHAC

Strong thermodynamic energy can stabilize the double-stranded structure between the primer and target sequence, ensuring amplification efficiency. However, it may also compromise primer specificity. Hence, we first investigated the performance of primers with various thermodynamic energies and optimized the design of thermodynamic parameters through theoretical modeling.

Based on the theoretical model that primarily considers primer hybridization and extension (Supplementary Section 3), we explored amplification efficiency and specificity under different primer free energies. To ensure the uniformity of amplification efficiency, the most direct way is to make all primers achieve high amplification efficiency (Fig. 3a). The results indicated a positive correlation between amplification efficiencies and the ΔG° of FP and RP while $\Delta G^\circ = -8 \text{ kcal mol}^{-1}$ is a boundary for at least half of primers binding to templates. More negative free energy implies a stronger binding capacity between the primer and the template, however, it may include partially complementary non-specific templates, reducing the primer's specificity. To quantitatively characterize this relationship, we further examined primer specificity at different ΔG° values (Fig. 3b). The results revealed that more positive free energy of primer hybridization led to higher specificity, contradicting the requirement for optimal amplification efficiency. To consider an appropriate trade-off between specificity and sensitivity, we calculated the odds ratio (OR) values under different primer free energies as a reference index. Optimal OR values were obtained when the primer's free energy was approximately $-11.5 \text{ kcal mol}^{-1}$ (Fig. 3c). A detailed comparison of OR values between FL- and FE-designs indicated that FE-primers exhibit higher OR values with a more concentrated distribution (Fig. 3d, e), further demonstrating the superiority of the FE-design.

Subsequently, we designed 12 primer sets for experimental studies: 6 sets of FE-primers and 6 sets of FL-primers, based on plasmid sequences at 6 different positions (Fig. 3f). For systematic and cost-effective exploration of the hybridization kinetics and thermodynamic yields of each primer, we employed the X-Probe architecture, leveraging universal oligonucleotides with fluorophores or quenchers to significantly reduce synthesis costs (Supplementary Fig. 19). Figure 3g shows the experimental and simulation results for one primer as an example, demonstrating good concordance between experimental and simulation results, affirming the reliability of our simulations (other data is shown in Supplementary Fig. 23 and Supplementary Fig. 24). According to the formula $\Delta G^\circ = -RT \ln K_{eq}$, when $A = B$, we

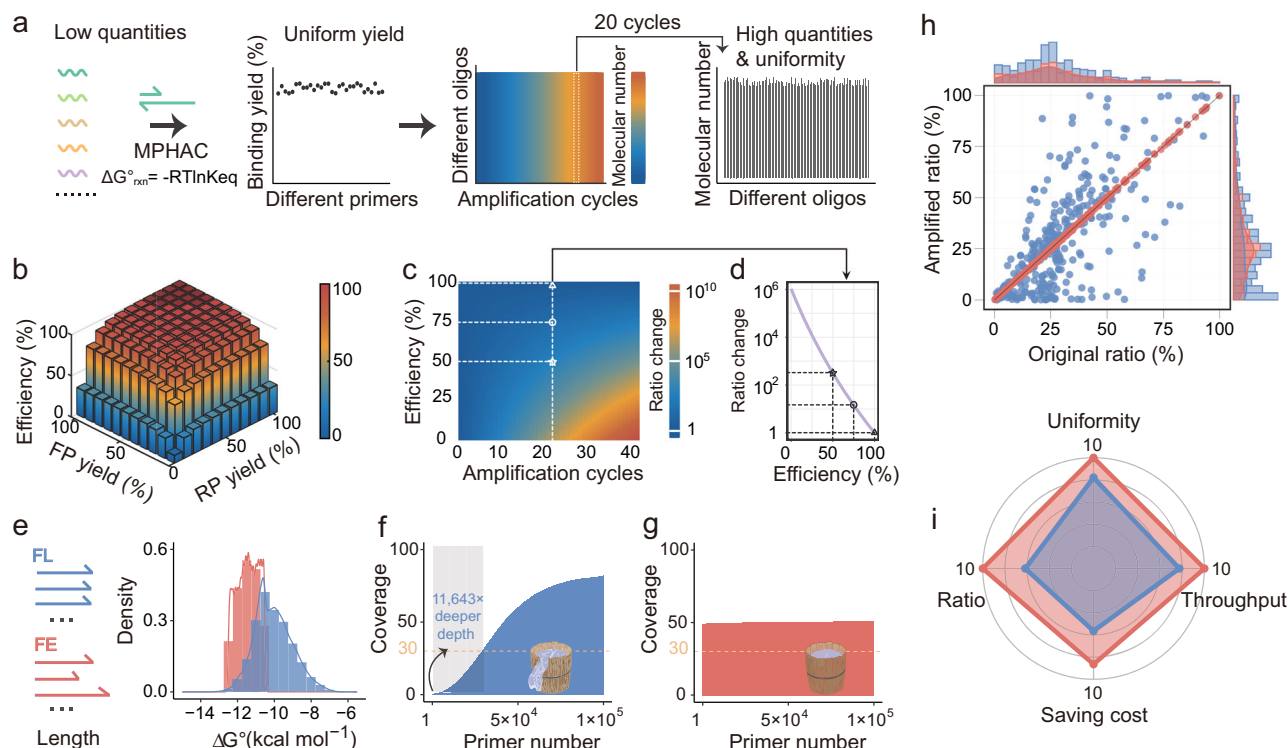


Fig. 2 | Simulation comparison between MPHAC and conventional fixed length primer amplification. **a** Employing MPHAC for selective amplification of microchip-synthesized DNA to enhance molecular quantity while maintaining sequence homogeneity. **b** Relationship between primer amplification efficiency and hybridization yield. **c** In the presence of two files of equal size, the variation in the file ratio before and after amplification (defined as ratio change) spans over 10 orders of magnitude, influenced by different PCR cycles and their respective amplification efficiencies. **d** Ratio change at various amplification efficiencies with 20 PCR cycles. Conventional pathway is often based on fixed-length (FL) primers, while MPHAC achieves consistent amplification efficiency by rigorously controlling the primers with fixed-energy (FE). For comparing conventional pathway with MPHAC, we randomly generated 100,000 primer pairs using both methods based on the human genome sequence, and simulated their performance. **e** Thermodynamic energy distribution of FL-primers and FE-primers. Analysis of

primer hybridization thermodynamic energy indicates that, even with a fixed length and restricted GC content, the free energy distribution of the primers remains wide, causing variations in their amplification efficiencies. **f** Simulated sequencing results of FL-primers under 50× sequencing depth. Using a 30× threshold, detecting all sequences requires an 11,643-fold deeper sequencing depth. **g** Simulated sequencing results of FE-primers under 50× sequencing depth. In contrast, MPHAC achieves consistent amplification efficiency by rigorously controlling the primer free energy distribution, ensuring the effective detection of all sequences. **h** Comparison of ratio consistency before and after amplification using FL-primers vs. FE-primers. R^2 is the square of Pearson's correlation coefficient, calculated from the total sum of squares and the residual sum of squares. **i** Performance comparison between FL-primers (blue) and FE-primers (red). The calculation for each metric is described in the "Methods section". Source data are provided as a Source Data file.

observed that the variation in hybridization yield within the FE-group was ~20%, with the yield distribution predominantly concentrated in the upper and lower 50%. Conversely, the disparity in the FL-group exceeded 80% (Fig. 3h, left). Given the typically low initial concentration of the template during actual PCR, we simulated the hybridization yield of primers when $A \gg B$. Results indicated that all FE-primers achieved high levels (>80%), while FL-primers ranged from 4.5% to 98.9% (Fig. 3h, right). Summarily, FE-primers exhibited a higher and more concentrated range of yields in both cases, whereas certain FL-primers showed negligible hybridization yields. This disparity directly influences subsequent PCR amplification efficiency. Higher yields signify a stronger affinity between the primer and template, enhancing the probability of successful enzyme binding with the template and facilitating the crucial extension step in the PCR process.

Finally, building upon the prior primer design, we explored the impact of thermodynamic control on PCR uniformity. The simulation results revealed that the amplification efficiency within the FE-group consistently ranged from 99.8% to 99.9%, whereas the FL-group exhibited a broader range from 36.6% to 99.7% (Supplementary Fig. 18). Using plasmid templates of 6000 copies, we observed remarkable homogeneity in the amplification efficiency of the FE-group, with all six primer sets displaying Ct values around 23.5. Conversely, the FL-group exhibited varying degrees of deviation, with

three primer sets demonstrating increased Ct values ranging from 1.2 to 19 (Fig. 3i).

Homogeneous amplification and lossless index of MPHAC-DIS

To further validate the effectiveness of MPHAC, we implemented it in DNA information storage (MPHAC-DIS) (Supplementary Fig. 25). Since all files are fundamentally binary data, during the DNA storage encoding phase, we converted 1 byte (8 bits) of binary information into 5 nucleotides based on a predefined transcoding table (Supplementary Fig. 26). Primer binding regions were then appended to both ends of the information sequence to enable random access (Supplementary Fig. 27). In MPHAC-DIS, 35,406 DNA oligos were synthesized to achieve the concurrent encoding and storage of multimodal data, encompassing texts, images, and videos. During the decoding process, the most plausible sequences were selected for information retrieval using frequency statistics (Supplementary Fig. 31). Supplementary Table 3 and Supplementary Table 4 showed the detailed file information and access combinations.

Here, we present the decoding results for one specific set of primer combinations, which accessed eight files simultaneously (Fig. 4a). The results revealed minimal regional variations in coverage distributions among the majority of files, and the fold-80 base penalties for each file remained below 1.7 (Supplementary Fig. 32), which indicated

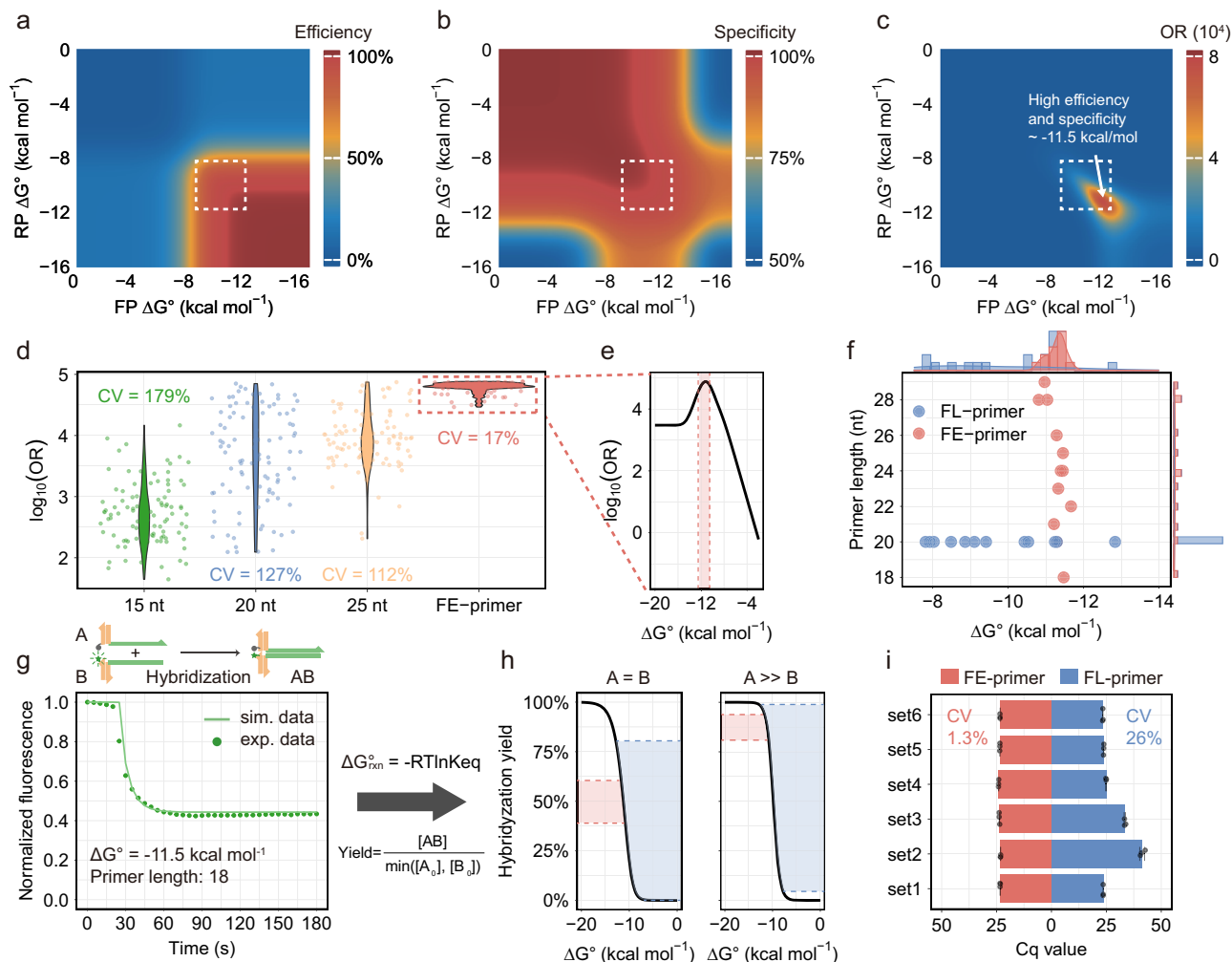


Fig. 3 | Parameters optimization of MPHAC. **a–c** Simulated results of amplification efficiency (**a**), specificity (**b**) and odds ratio (OR) (**c**) under different primer free energies based on the theoretical model. Details of the model construction are provided in Supplementary Section 3. The white dashed box represents the free energy range for 80% of FL-primers in 100,000-plex primer pairs. Amplification efficiency positively correlates with free energy, while specificity exhibits the opposite trend. We considered the amplification efficiency and specificity together by OR, and the results suggest that the primer free energy can be optimized around $-11.5 \text{ kcal mol}^{-1}$. **(d)**, **e** OR values under different length FL-primers and FE-primers (**d**) and different primer free energies (**e**). Here, the coefficient of variation (CV) is a measure of the relative dispersion of a data set, expressed as the ratio of the standard deviation to the mean. **f** Primer length and free energy distribution. Six sets of FE-primers and six sets of FL-primers were designed based on plasmid

sequences. **g** Monitoring primer hybridization fluorescence kinetics using X-probe. Experimental data are represented by dots, and simulated data is depicted by lines. **h** Simulated hybridization yields for primers when $A = B$ (left) and $A \gg B$ (right). Here, A refers to the primer, and B denotes the reverse complementary sequence of the primer (template). The range of hybridization yields for FE-primers is considerably narrower (pink) than that of FL-primers (blue). **i** PCR results for FE- and FL-primers. Data are presented as mean \pm standard deviation (SD), based on measurements from 3 independent samples. Here, Ct value (Cycle threshold value) represents the cycle number at which the fluorescence signal exceeds a defined threshold, indicating the presence of the target. Theoretically, identical concentrations of the target should result in the same Ct value. Source data are provided as a Source Data file.

good amplification uniformity within the files based on MPHAC. For inter-file homogeneity, a consistent level of coverage across all files was observed under the FE-design, while more than an order of magnitude of variation exists under the FL-design. (Fig. 4b). Compared to the FL-design, the fold-80 base penalty for the FE-design was reduced from 6.0 to 1.4, indicating that MPHAC effectively maintained amplification homogeneity across the files (Fig. 4c). By analyzing the ratio of each file in the sequencing results of multiple experiments, we observed that, under the FE-design, the experimental ratios closely aligned with the expected ratios, displaying a linear correlation in distributions, with the correlation coefficient r^2 increasing from 0.91 to 0.96 compared to the FL-design (Fig. 4d). This further indicated that MPHAC effectively maintains excellent amplification uniformity across the files. Comparing the decoding results under the two designs, the average correct rate for most files under the FE-design was close to

100%, while the FL design showed varying degrees of decrease, with the lowest being only 8% (Fig. 4e). Repeated experiments with preservation under different buffers showed that MPHAC-DIS also demonstrated good reproducibility and robustness (Supplementary Figs. 34–39).

We also conducted a detailed error analysis of MPHAC-DIS (Supplementary Section 8). In the sequencing results, $\sim 80\%$ of the reads were completely correct, while errors existed in 20% of the reads (Supplementary Fig. 35). Among the erroneous sequences, some appeared in multiple experiments (Supplementary Fig. 40). Additionally, within the error sequences, certain sequences shared the same payloads (Supplementary Figs. 44–46), and most of these payloads had relatively high GC content ($> 55\%$, Supplementary Fig. 41). These sequences often formed secondary structures either within the payloads themselves or with the primer binding regions (Supplementary

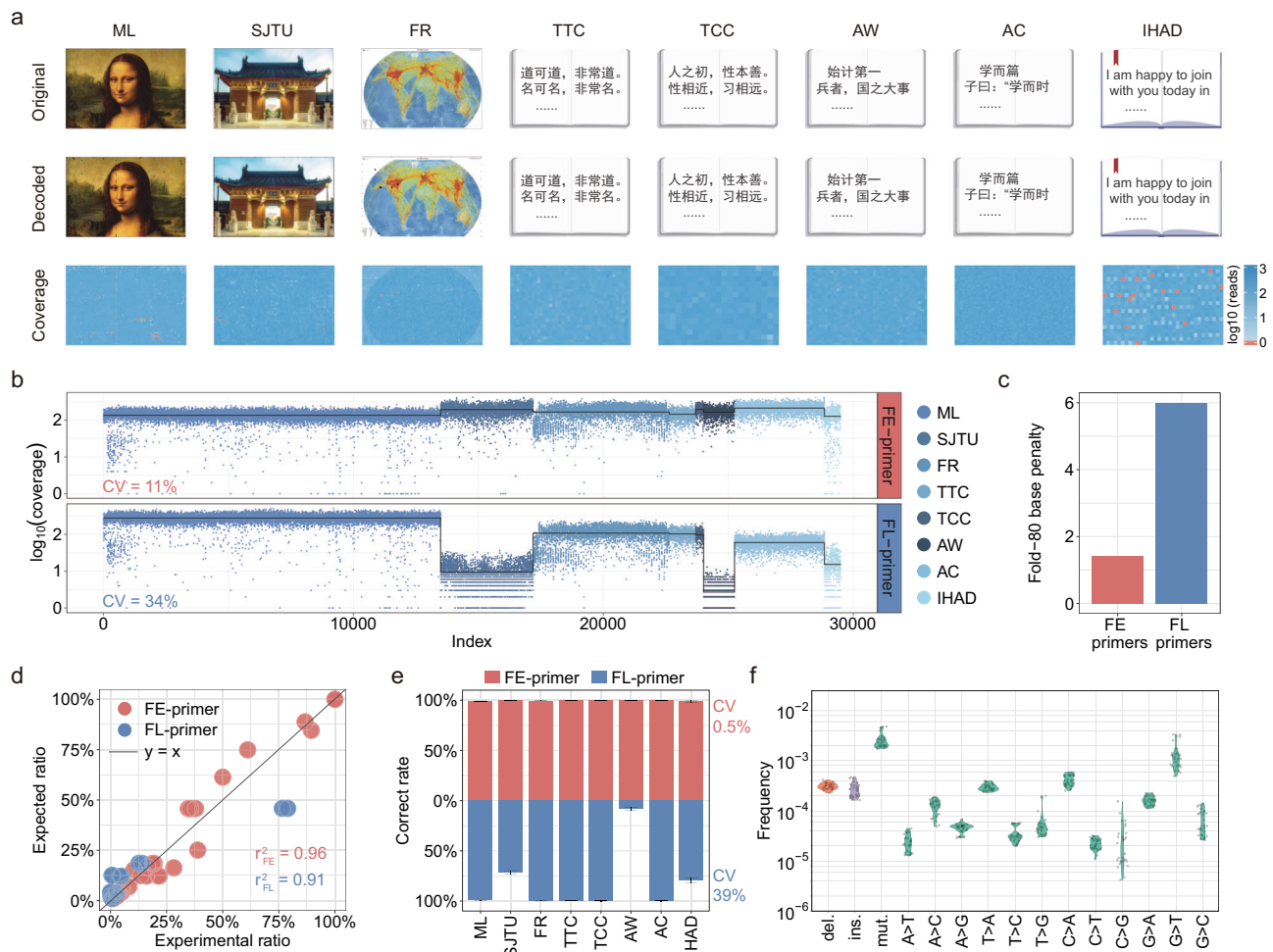


Fig. 4 | Homogeneous amplification and lossless index of MPHAC-DIS. We encoded files of multimodal data, including text, pictures, etc., based on the encoding principle (Supplementary Fig. 26). **a** Simultaneous retrieval of 8 distinct files in the DNA pool using primer combinations. The top-to-bottom presentation includes original files, decoding results, and the distribution of sequence coverage, with red pixels indicating repeat reads lower than 5. Notably, the sequence coverage within each file is uniformly distributed, showcasing excellent intra-file homogeneity. **b** Coverage comparison between FE-primers and FL-primers. The results illustrate a high degree of consistency and highlighting the outstanding inter-file amplification homogeneity of MPHAC-DIS. **c** Fold-80 base penalty of FE-primers and FL-primers. Fold-80 base penalty is a metric used to assess coverage

uniformity in sequencing experiments. It indicates how much more sequencing is needed for 80% of the targets to reach the mean coverage. The closer the Fold-80 base penalty is to 1, the better the coverage uniformity. **d** Correlation between the experimental ratios and the expected ratios of various files. Through multiple random-access experiments, we observed a linear relationship between the actual and expected ratio, providing additional confirmation of the inter-file amplification homogeneity of MPHAC-DIS. **e** Decoding accuracy comparison between FE primers and FL primers. The correct rate for the majority of files exceeds 99%, underscoring the exceptional performance of MPHAC-DIS. **f** Statistics on the frequency of sequence error types in MPHAC-DIS. Data points were obtained from the results of 30 libraries. Source data are provided as a Source Data file.

Figs. 47–49 and Supplementary Fig. 51). Further analysis of the relationship between errors and base positions revealed that the likelihood of errors increased toward the 3' end of the sequencing reads (Supplementary Figs. 52 and 53). This may be related to the higher error rates associated with DNA synthesis and sequencing at longer lengths. Here, it's noted that sequencing reads are reverse complements of the synthesized sequences, errors introduced at the 5' end of the synthesized DNA appear at the 3' end of the sequencing reads. When analyzing the types of errors, we found that single-base substitutions were the most common, with G > T substitutions being the most frequent (Fig. 4f). Our results are consistent with Erlich's findings, and quantifying errors helps in understanding their sources as well as correcting them^{12,34}.

Study of MPHAC-DIS using ultra-low coverage sequencing

Sequencing depth directly affects data reliability and precision. Higher sequencing depth improves error correction and data confidence.

However, there is a trade-off between sequencing depth and cost due to increased resource and time demands. Therefore, evaluating the operational efficiency of MPHAC-DIS at various sequencing depths is crucial to determining the optimal depth for the intended decoding purposes.

We acquired sequencing results of various depths, ranging from 550× to 4×, through gradient dilution prior to sequencing (Fig. 5a). Comparing to FL-design, the density distribution of correct reads across libraries always exhibited a narrow range under FE-design, demonstrating its superior amplification homogeneity (Fig. 5b). The decoding results indicated that at a sequencing depth of 17×, the average correct rate under FE-design remained above 90%, while it was only 54.5% under FL-design (Fig. 5c). To further validate our results, we performed down-sampling on the previously acquired sequencing data, adjusting sample sizes in accordance with the experimental conditions. The results demonstrated a robust linear relationship between the down-sampling data and the experimental sequencing

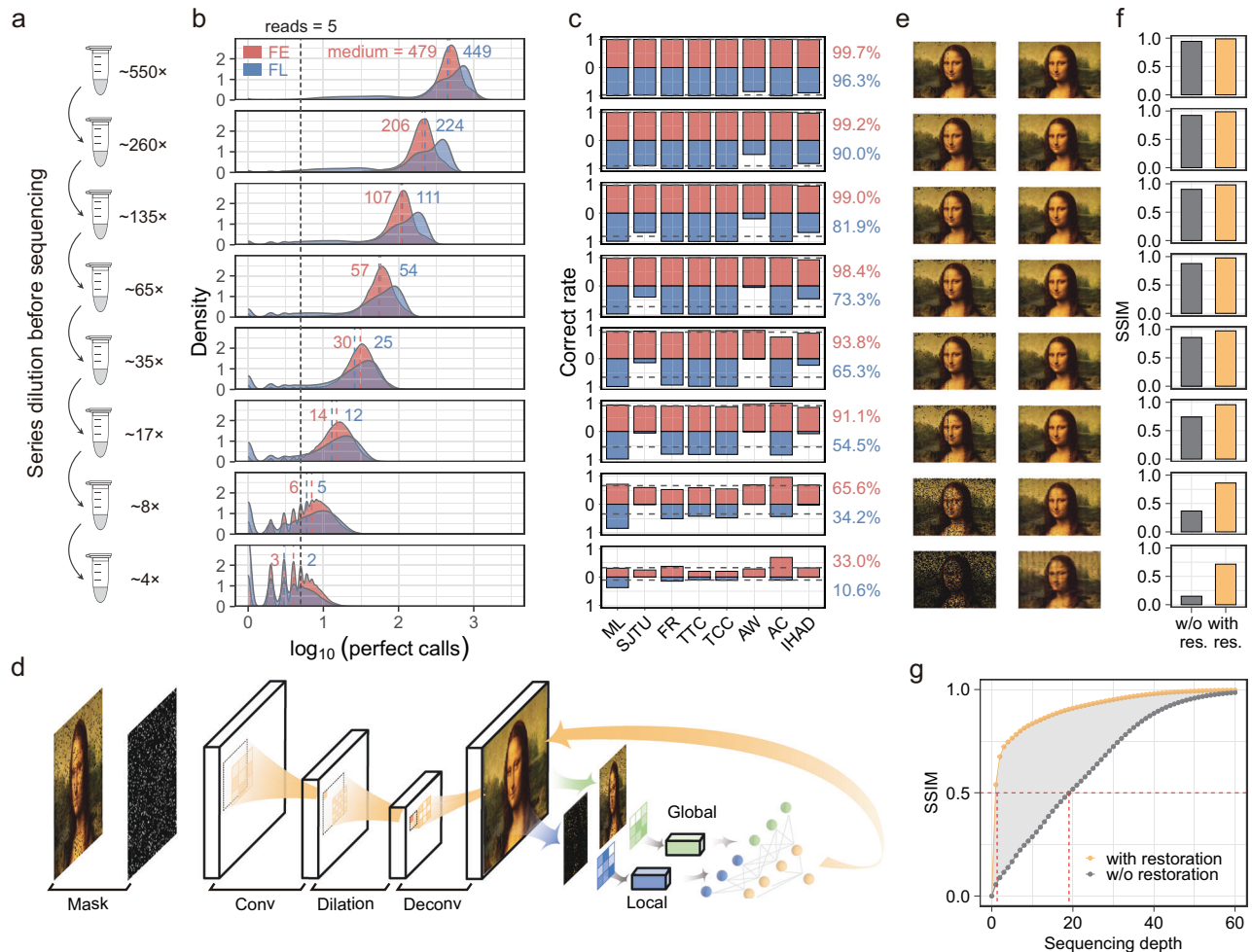


Fig. 5 | Study of MPHAC-DIS using ultra-low coverage sequencing. a The amplified libraries were serially diluted before sequencing, and then mixed together for sequencing. Finally, libraries with different sequencing depths were obtained by demultiplexing the sequencing results. **b** Density distribution of the exact matching sequences. The red and blue dashed lines indicate the median number of perfect calls under the FE-design and the FL-design, respectively. **c** Decoding accuracy for each library file. The red and blue dashed lines indicate the average decoding correct rate for the FE-design versus the FL-design, respectively. **d** Image restoration strategy employing a convolutional neural network. Detailed information can be seen in Supplementary Fig. 57. **e** Decoding results of ML image without restoration (left) and the results with restoration (right). **f** Comparison of the

corresponding structural similarity index measure (SSIM, a metric used to measure the similarity between two images) of ML image before (w/o res.) and after restoration (with res.). The SSIM values indicate the similarity of each image to the original image. The result reveals that the image restoration technique has a more pronounced impact at lower sequencing depths. **g** Fitting results for the quality of decoded images with and without image restoration at lower sequencing depths. Using SSIM > 0.5 as the threshold, a sequencing depth of 19× is necessary before image restoration. However, with image restoration, a sequencing depth of only 1× is needed, showcasing the efficacy of image restoration technology in further reducing sequencing costs. Source data are provided as a Source Data file.

data ($r^2 = 0.97$), providing additional confirmation that MPHAC-DIS can effectively reach the theoretical limit (Supplementary Fig. 56).

Benefiting from consistent within-file amplification uniformity, decoding errors (black dots) were evenly distributed across the image. This characteristic implies that each missing piece of information is surrounded by correct data to varying degrees, contributing to enhanced data recovery. Consequently, we developed a generative adversarial network for image restoration on the decoded images (Figs. 5d and Supplementary Fig. 57). The restoration results showed that even images with severe defects exhibited significant improvement, visible to the naked eye, attributable to the homogeneous distribution of errors (Fig. 5e). To more accurately quantify the impact of the image restoration technique, we introduced the Structural Similarity Index Measure (SSIM) to assess the image quality. The results indicated that the image restoration technique effectively enhances image quality, particularly when the sequencing depth

is below 30× (Fig. 5f). Fitting the decoding performance of MPHAC-DIS at low sequencing depths based on the available data revealed that even in the extreme case with a sequencing depth of only 1×, the combination with the image restoration technique still achieved SSIM > 0.5 (~80% correct decoding rate, Fig. 5g).

Environmental factors are also critical to the long-term stability of DNA data storage systems. Temperature fluctuations, for instance, can accelerate DNA degradation, with higher temperatures exacerbating hydrolytic and oxidative damage to DNA strands. Humidity also plays a significant role in DNA integrity, as excessive moisture introduces water molecules that promote hydrolysis, ultimately leading to data loss. Furthermore, extended exposure to UV light can trigger thymine dimer formation and other photochemical reactions, further compromising the stored DNA. Studies have demonstrated that dry storage, particularly when using encapsulation techniques, can effectively mitigate these risks by creating a stable, inert environment that shields

DNA from external factors. Additionally, for DNA pools where some damage has occurred, data restoration techniques offer another solution. The combination of preservation strategies with restoration methods helps to ensure a more robust DNA storage system.

Discussion

Our MPHAC method achieves selective, high-throughput and homogeneous amplification for DNA templates while retaining the original template concentration ratio information. Differing from FL-primers, MPHAC can achieve superior homogeneity with a larger primer pool. Its non-discriminatory and selective amplification relies on precisely regulated primer amplification based on ΔG° , with the introduction of OR values considering both efficiency and specificity. To predict and regulate amplification efficiency based on sequences, we constructed a two-step model: a hybridization step and an extension step. This generalized model enables high-throughput prediction of amplification efficiency based on primer and template sequences.

The MPHAC-DIS achieves random access with high-fidelity reading of files for high-throughput storage purposes. DNA information storage necessitates alignment with real application scenarios, wherein the number of preserved files surpasses that of the mere dozens retained in current traditions. As the number of saved files escalates, homogeneous amplification becomes paramount. Homogeneous amplification significantly curtails read costs, particularly as the number of files surges by orders of magnitude, as evidenced by a reduction in sequencing depth. This reduction in sequencing depth, enabled by homogeneous amplification and coupled with AI-driven information recovery techniques, facilitates information decoding at shallower depths. Moreover, the MPHAC-DIS method ensures lossless preservation of information, with homogeneous amplification maintaining the original template proportions, thus enhancing fidelity in information preservation.

While the MPHAC-DIS method is effective for DNA information storage, it does have certain limitations. Firstly, although we optimized and generated a library comprising tens of thousands of primer pairs through our primer design workflow, practical wet-lab experiments face challenges in handling such a large number simultaneously, which can lead to non-specific amplification and primer dimer formation. Beyond algorithmic solutions, primer dimers and longer non-specific amplicons can be reduced using size selection with magnetic beads and enzymatic cleavage. Secondly, even tens of thousands of primer pairs are insufficient for large-scale DNA storage. Future advancements will require faster algorithms and methods capable of generating primers with higher capacity. Additionally, techniques such as isothermal amplification can be employed to improve access accuracy during random access, and physical addressing remains a viable option. Thirdly, the potential occurrence of partially complementary sequences within the oligonucleotide outside the primer-binding region could also impact the uniformity of amplification. To address this, more flexible encoding strategies can be employed to avoid motifs related to primers and their complementary sequences within the payload. Finally, biases introduced during the DNA synthesis stage can affect the reliability of the storage system. To mitigate this, we can introduce error correction codes and incorporate partial redundancy to ensure decoding accuracy. Moreover, there is a need to develop more economical and higher-quality DNA synthesis techniques, such as enzymatic synthesis.

We believe the programmability and selectivity of MPHAC can be extended to other areas. First, MPHAC can predict the efficiency of amplification through precise regulation of the hybridization process, enabling control over the hybridization efficiency of programmable primers and, in turn, the amplification efficiency for different sequences. Second, the selective programmable

amplification of templates can be used in preview systems for constructing files in DNA information storage. Moreover, beyond DNA information storage, MPHAC can be used for genome target sequence enrichment and disease diagnosis, facilitating specific amplification of various targets.

Methods

Primer generation

First, 2,000,000 primers of 15–30 nt were randomly generated. Subsequently, the standard free energy of each primer was calculated using the nearest neighbor model and primers with free energy in the range of $-10.5 \text{ kcal mol}^{-1}$ to $-12.5 \text{ kcal mol}^{-1}$ were retained. The GC content check (45%–55%) and homopolymer check (AAAA, TTTT, CCC, GGG) of the primers were then performed. For retained primers, badness (a metric describing the probability of dimer formation between two primers) between all primers was calculated. Dimer check was performed using max badness, and only primers with max badness less than 1 were retained. Furthermore, the minimum free energy (MFE) structure of each primer was analyzed by NUPACK. Primers with free energy equal to 0 kcal mol^{-1} passed the secondary structure check and were included in the final primer library. Finally, crosstalk between primers and between primers and their reverse complementary sequences was verified using BLAST to ensure that the number of complementary bases (including shift complementarity) was less than or equal to 16.

Theoretical simulation for PCR with different primers

To simplify the model, the hybridization between primers and templates and enzyme-assisted primer extension in a single PCR cycle were mainly considered. In contrast, factors such as enzyme activity changes, interactions between enzyme and double-stranded DNA, and denaturation efficiency were not considered in most situations. A series of ordinary differential equations were constructed based on the reactions occurring in a single PCR cycle. These ordinary differential equations were simulated using MATLAB's stiff "ode23s" solver, with a relative tolerance of 10^{-6} and an absolute tolerance of 10^{-30} . The final reactant concentration (typically the result at 30 s) was used to calculate PCR amplification efficiency. More detailed information is provided in Supplementary Section 3.

100,000-plex primer simulation

Firstly, 100,000 primers were generated for each of the FL-design and FE-design, respectively. The ΔG° of each sequence was then calculated to determine the free energy distributions for both designs. Subsequently, the amplification efficiency of each primer was calculated based on the theoretical model, and the coverage of each primer was derived using the following equation:

$$Coverage_i = \frac{Depth \cdot n \cdot Efficiency_i^{Cycle}}{\sum_{i=1}^n Efficiency_i^{Cycle}} \quad (1)$$

Here, n represents the total number of primers, which in our case is 100,000. Depth refers to the sequencing depth, set to $50\times$, and Cycle denotes the number of PCR amplification cycles, set to 20.

Calculation of metrics

In Fig. 2i, we use four different metrics to compare the FE-design with the FL-design: uniformity, ratio, saving cost and throughput. Uniformity corresponds to the fold-80 value of the simulated sequencing results, ratio corresponds to the Pearson's correlation coefficient r^2 between the file ratios before and after amplification, saving cost corresponds to the sequencing depth, and throughput corresponds to the primer capacity. Additionally, we normalized all the metrics using

the following formulas:

$$Uniformity_{norm} = 10 - (Uniformity_{original} - 1) \quad (2)$$

$$Ratio_{norm} = Ratio_{original} \bullet 10 \quad (3)$$

$$Cost_{norm} = 10 - \log_{10}(Cost_{original}) \quad (4)$$

$$Throughput_{norm} = 10 - \log_{10}\left(\frac{Throughput_{original}}{Throughput_{max}}\right) \quad (5)$$

Here, X_{norm} denotes the normalized value of the metric, while $X_{original}$ denotes the initial value.

In Fig. 3, we explored the amplification efficiency, specificity, and OR under different primer free energies. The formulas for the relevant metrics are as follows:

$$Efficiency = \frac{Conc_{amp}}{Conc_{ori}} - 1 \quad (6)$$

$$Specificity = \frac{Efficiency_{target}}{Efficiency_{target} + Efficiency_{non-target}} \quad (7)$$

$$OR = \frac{Efficiency / (1 - Efficiency)}{(1 - Specificity) / Specificity} \quad (8)$$

$Conc_{ori}$ represents the initial concentration of the target sequence, while $Conc_{amp}$ denotes the concentration of the target sequence after amplification. $Efficiency_{target}$ signifies the amplification efficiency of the matching target sequence, and $Efficiency_{non-target}$ indicates the amplification efficiency of the non-target. Here, non-target denotes a non-specific template with multiple base differences compared with target, whose ΔG with the primer is more positive than that of target with primer, and the difference between the two is $\Delta\Delta G$, which is fixed.

Oligonucleotides ordering

Oligo pools were synthesized by Twist Biosciences (San Francisco, USA) and delivered in the form of DNA powder. All DNA oligonucleotides used in this study, except for oligo pools, were purchased from Sangon Biotech Co. (Shanghai, China). Unmodified DNA oligonucleotides were purified via high affinity purification (HAP), and modified DNA oligonucleotides were purified via high-performance liquid chromatography (HPLC).

Time-based fluorescence acquisition

Prior to fluorescence kinetic monitoring, the fluorescent/quenching strand is pre-incubated with the corresponding strands, allowing their ends to be labeled with a fluorophore/quencher. The corresponding oligonucleotides were annealed in a polymerase chain reaction (PCR) thermal cycler. The temperature was set at 95 °C for 5 minutes initially and then decreased to 25 °C over 2 hours. Time-based fluorescence data corresponding to the hybridization reactions were then acquired using a Horiba Fluoromax instrument. The excitation and emission wavelengths for the FAM fluorophore used on the X-Probe were optimized to 494 nm and 518 nm, respectively.

Amplification homogeneity characterization via qPCR

The qPCR assays were performed on a CFX96 Touch Real-Time PCR Detection System using 96-well plates (Bio-Rad). Blue SYBR Green Master Mix (YEASEN, cat. no. 11184ES03) was used for enzymatic amplification and fluorescence signal generation. In a typical PCR protocol, 5 μ L of Blue 2 \times Master Mix, 1 μ L of DNA template (6000

copies/ μ L), 1 μ L of forward primers (4 μ M), 1 μ L of reverse primers (4 μ M), and the appropriate volume of ddH₂O were mixed to a final volume of 10 μ L. Thermal cycling started with a 3-minute incubation step at 95 °C, followed by 40 cycles of 10 seconds at 95 °C for DNA denaturing and 30 seconds at 60 °C for annealing and extension.

Data encoding and decoding

Multi-modal files including text, images, and videos were stored in DNA. All files were first segmented into small pieces of information and assigned address information. Subsequently, they were transformed into DNA sequences based on a coding strategy. Finally, primer sequences were added to the ends, and then the oligo pool was ordered from the DNA synthesis company. For decoding, the reads were first aligned with the primer sequences to determine their acceptability and the corresponding files. Then, the address and payload information in the reads were extracted, and the payload of each read was stored based on the address information. The payloads with the highest frequency of occurrences and greater than a threshold value under each address were used to decode and recover the stored data. More detailed information is provided in Supplementary Section 5.

Visualization of Sequencing Coverage Results

To visually represent the sequencing coverage results, we employed the following approach. During the encoding phase, the information is split and stored across multiple sequences. After sequencing, the coverage of each sequence is calculated by aligning the reads, resulting in an $N \times 1$ vector, where N represents the number of sequences. To transform this vector into a two-dimensional format, it is rearranged into an $M \times K$ matrix, with M and K chosen to suit the visualization needs. For image files, this vector is rearranged to match the structure of the original image, allowing the coverage data to align visually with the image layout. Finally, a heatmap is generated from this matrix, with each point in the heatmap representing the coverage of a specific sequence, thereby providing a clear and intuitive visual representation.

Files random access and NGS sequencing library preparation

First, primers corresponding to the desired files were mixed to a final concentration of 4 μ M. DNA sequences were amplified using Phusion DNA polymerase (Thermo Fisher, cat. no. F-530L). The reaction mixture (50 μ L) consisted of 2 μ L oligo pool, 2.5 μ L forward primer mix, 2.5 μ L reverse primer mix, 0.5 μ L polymerase, 1 μ L dNTP (10 mM), 10 μ L 5 \times HF buffer, and 31.5 μ L ddH₂O. The PCR conditions included an initial denaturation step at 98 °C for 2 minutes, followed by 12 cycles of 98 °C for 20 seconds, 63 °C for 30 seconds, and 72 °C for 30 seconds. Amplicon products were purified from the reaction using a magnetic bead-based DNA cleanup kit (Vazyme, cat. no. N411-02).

In the next step, sequencing adapters were appended. For this, 15 μ L of the purified amplicon was mixed with 2.5 μ L adapter forward primer mix, 2.5 μ L adapter reverse primer mix, 0.5 μ L polymerase, 1 μ L dNTP (10 mM), 10 μ L 5 \times HF buffer, and 18.5 μ L ddH₂O, to a total volume of 50 μ L. The reaction was thermocycled as follows: an initial denaturation step at 98 °C for 2 minutes, followed by 3 cycles of 98 °C for 20 seconds, 63 °C for 30 seconds, and 72 °C for 30 seconds. Products were once again purified from the reaction mixture to remove residual reagents.

Subsequently, the purified amplicons were diluted 100-fold and served as templates for a standard qPCR assay. A 10 μ L reaction consisted of 5 μ L Blue 2 \times Master Mix, 3 μ L diluted amplicons, 1 μ L N5 primer (diluted 5 \times), and 1 μ L N7 primer (diluted 5 \times). The thermocycling protocol included an initial step at 95 °C for 3 minutes, followed by 40 cycles of 10 seconds at 95 °C (denaturation) and 30 seconds at 60 °C (annealing and extension). The qPCR results were analyzed to determine the Ct values for each sample.

Finally, index sequences were attached to the amplicons using an indexing kit (YEASEN, cat. no. 12416ES96). The reaction mix (50 μ L)

included 15 μL of diluted amplicons, 1 μL N5 primer, 1 μL N7 primer, 0.5 μL polymerase, 1 μL dNTP (10 mM), 10 μL 5 \times HF buffer, and 21.5 μL ddH₂O. The thermocycling program consisted of an initial denaturation at 98 °C for 2 minutes, followed by Ct+4 cycles of 98 °C for 20 seconds, 63 °C for 30 seconds, and 72 °C for 30 seconds. After purification, the indexed amplicons were quantified using a quantification kit (Sangon, cat. no. N608301-0500). Based on the quantification results, all libraries were pooled and prepared for NGS sequencing.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available in the manuscript or the Supplementary Information. The experimental sequencing data have been deposited in the European Nucleotide Archive under accession code [PRJEB82436](https://www.ebi.ac.uk/ena/record/PRJEB82436). Source data are provided with this paper.

Code availability

The codes used in this study have been deposited in GitHub and are accessible at <https://github.com/NABMELab/MPHAC-DIS>³⁵. The sequencing data, figure raw data, and a decoding demo have been deposited in Figshare at https://figshare.com/projects/MPHAC-DIS_Seq/217696.

References

- Extance, A. How DNA could store all the world's data. *Nature* **537**, 22–24 (2016).
- Ceze, L., Nivala, J. & Strauss, K. Molecular digital data storage using DNA. *Nat. Rev. Genet.* **20**, 456–466 (2019).
- Organick, L. et al. Probing the physical limits of reliable DNA data retrieval. *Nat. Commun.* **11**, 616 (2020).
- Meiser, L. C. et al. Synthetic DNA applications in information technology. *Nat. Commun.* **13**, 352 (2022).
- Nguyen, B. H. et al. Scaling DNA data storage with nanoscale electrode wells. *Sci. Adv.* **7**, eabi6714 (2021).
- Ma, Y., Zhang, Z., Jia, B. & Yuan, Y. Automated high-throughput DNA synthesis and assembly. *Heliyon* **10**, e26967 (2024).
- Hoose, A., Vellacott, R., Storch, M., Freemont, P. S. & Ryadnov, M. G. DNA synthesis technologies to close the gene writing gap. *Nat. Rev. Chem.* **7**, 144–161 (2023).
- Hughes, R. A. & Ellington, A. D. Synthetic DNA synthesis and assembly: putting the synthetic in synthetic biology. *Cold Spring Harb. Perspect. Biol.* **9**, a023812 (2017).
- Tian, J. et al. Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* **432**, 1050–1054 (2004).
- Linnarsson, S. Recent advances in DNA sequencing methods – general principles of sample preparation. *Exp. Cell Res.* **316**, 1339–1343 (2010).
- Gao, Y., Chen, X., Qiao, H., Ke, Y. & Qi, H. Low-bias manipulation of DNA oligo pool for robust data storage. *ACS Synth. Biol.* **9**, 3344–3352 (2020).
- Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950–954 (2017).
- Koch, J. et al. A DNA-of-things storage architecture to create materials with embedded memory. *Nat. Biotechnol.* **38**, 39–43 (2020).
- Gimpel, A. L., Stark, W. J., Heckel, R. & Grass, R. N. A digital twin for DNA data storage based on comprehensive quantification of errors and biases. *Nat. Commun.* **14**, 6026 (2023).
- Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. *Science* **337**, 1628 (2012).
- Organick, L. et al. Random access in large-scale DNA data storage. *Nat. Biotechnol.* **36**, 242–248 (2018).
- Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77–80 (2013).
- Liu, Y. et al. Optimized PCR-based enrichment improves coverage uniformity and mutation detection in mitochondrial DNA next-generation sequencing. *J. Mol. Diagn.* **22**, 503–512 (2020).
- Mamanova, L. et al. Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7**, 111–118 (2010).
- Terekhov, S. S. et al. Liquid drop of DNA libraries reveals total genome information. *Proc. Natl Acad. Sci. USA* **117**, 27300–27306 (2020).
- Bögels, B. W. A. et al. DNA storage in thermoresponsive microcapsules for repeated random multiplexed data access. *Nat. Nanotechnol.* **18**, 912–921 (2023).
- Tawfik, D. S. et al. Amplification of complex gene libraries by emulsion PCR. *Nat. Methods* **3**, 545–550 (2006).
- Wu, L. R. et al. Continuously tunable nucleic acid hybridization probes. *Nat. Methods* **12**, 1191–1196 (2015).
- Zhang, D. Y., Chen, S. X. & Yin, P. Optimizing the specificity of nucleic acid hybridization. *Nat. Chem.* **4**, 208–214 (2012).
- Wang, J. S. & Zhang, D. Y. Simulation-guided DNA probe design for consistently ultraspecific hybridization. *Nat. Chem.* **7**, 545–553 (2015).
- Wu, Y. et al. Enhanced detection of novel low-frequency gene fusions via high-yield ligation and multiplexed enrichment sequencing. *Angew. Chem. Int. Ed.* **63**, e202316484 (2024).
- Shen, H. et al. Random Sanitization in DNA Information Storage Using CRISPR-Cas12a. *J. Am. Chem. Soc.* **146**, 35155–35164 (2024).
- Ping, Z. et al. Towards practical and robust DNA-based data archiving using the yin–yang codec system. *Nat. Comput. Sci.* **2**, 234–242 (2022).
- Grass, R. N., Heckel, R., Puddu, M., Paunescu, D. & Stark, W. J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew. Chem. Int. Ed.* **54**, 2552–2555 (2015).
- Xie, N. G. et al. Designing highly multiplex PCR primer sets with simulated annealing design using dimer likelihood estimation (SADDLE). *Nat. Commun.* **13**, 1881 (2022).
- Ouldrige, T. E., Šulc, P., Romano, F., Doye, J. P. K. & Louis, A. A. DNA hybridization kinetics: zippering, internal displacement and sequence dependence. *Nucleic Acids Res.* **41**, 8886–8895 (2013).
- Stolovitzky, G. & Cecchi, G. Efficiency of DNA replication in the polymerase chain reaction. *Proc. Natl Acad. Sci. USA* **93**, 12947–12952 (1996).
- Sundaraman, B. et al. A method to generate capture baits for targeted sequencing. *Nucleic Acids Res.* **51**, e69 (2023).
- Heckel, R., Mikutis, G. & Grass, R. N. A characterization of the DNA data storage channel. *Sci. Rep.* **9**, 9663 (2019).
- Weng, Z. et al. Massively parallel homogeneous amplification of chip-scale DNA for DNA information storage (MPHAC-DIS). *GitHub* <https://doi.org/10.5281/zenodo.14068312> (2024).

Acknowledgements

This work was supported by the National Key R&D Program of China (No. 2022YFF1201800), the National Natural Science Foundation of China (Nos. 22174094, 22025404, T2188102), the Fundamental Research Funds for the Central Universities (YG2023QNA33), the Shanghai Science and Technology Committee (24Y22800300). Young Leading Scientists Cultivation Plan supported by Shanghai Municipal Education Commission (ZXWH1082101).

Author contributions

C.F. and P.S. conceived the research. P.S. and Z.W. designed the sequences. J.L. and Z.W. performed the majority of the experiments.

Y.W. constructed the data recovery algorithm. X.X., Z.W., and P.S. wrote the data analysis scripts. P.S. and Z.W. performed the simulations. P.S., C.F., X.Z., and Z.W. analyzed data and wrote the manuscript. F.W. and X.Z. participated in data analysis and discussions. C.F., P.S., Z.W., J.L., Y.W., X.X., F.W. and X.Z. reviewed and edited the manuscript.

Competing interests

P.S. and Z.W. declare the patent CN116959573A, and C.F., P.S., and Z.W. declare the patent CN116417071A, both of which are related to the primer design methods used in this work. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-55986-9>.

Correspondence and requests for materials should be addressed to Xiaolei Zuo, Ping Song or Chunhai Fan.

Peer review information *Nature Communications* thanks Jiongyu Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025