Parallel molecular data storage by printing epigenetic bits on DNA

https://doi.org/10.1038/s41586-024-08040-5

Received: 8 November 2023

Accepted: 11 September 2024

Published online: 23 October 2024

Open access

Check for updates

Cheng Zhang[1,8]✉, Ranfeng Wu[1,8], Fajia Sun[2,8], Yisheng Lin[1,8], Yuan Liang[1,3], Jiongjiong Teng[3], Na Liu[4,5], Qi Ouyang[2]✉, Long Qian[2]✉ & Hao Yan[6,7]✉

DNA storage has shown potential to transcend current silicon-based data storage technologies in storage density, longevity and energy consumption[1–3]. However, writing large-scale data directly into DNA sequences by de novo synthesis remains uneconomical in time and cost[4]. We present an alternative, parallel strategy that enables the writing of arbitrary data on DNA using premade nucleic acids. Through self-assembly guided enzymatic methylation, epigenetic modifications, as information bits, can be introduced precisely onto universal DNA templates to enact molecular movable-type printing. By programming with a finite set of 700 DNA movable types and five templates, we achieved the synthesis-free writing of approximately 275,000 bits on an automated platform with 350 bits written per reaction. The data encoded in complex epigenetic patterns were retrieved high-throughput by nanopore sequencing, and algorithms were developed to finely resolve 240 modification patterns per sequencing reaction. With the epigenetic information bits framework, distributed and bespoke DNA storage was implemented by 60 volunteers lacking professional biolab experience. Our framework presents a new modality of DNA data storage that is parallel, programmable, stable and scalable. Such an unconventional modality opens up avenues towards practical data storage and dual-mode data functions in biomolecular systems.

The markedly expanding global data-sphere has posed an imminent challenge on large-scale data storage and an urgent need for better storage materials[5,6]. Inspired by the way genetic information is preserved in nature, DNA has been recently considered a promising biomaterial for digital data storage owing to its extraordinary storage density and durability[1–3]. In current DNA storage, data is typically transcoded into nucleotide base sequences, and writing depends on de novo synthesis in which nucleotides are added one-by-one in predetermined orders[7]. Although de novo synthesis technologies have advanced continuously in throughput and efficiency[4,8], the serial synthesis process essentially limits the writing speed and the length of synthesized DNA, and prevents a substantial cost reduction in data writing[8,9] (Supplementary Fig. 1).

To enable efficient DNA storage, alternative data writing approaches that do not rely on de novo synthesis should work in both parallel and programmable fashions. In fact, some elegant designs have been proposed for parallel writing, for example, by structural programming of DNA carriers[10–12], but issues such as structural instability and limited throughput of information processing have impeded their application. In contrast, in human cells, the epigenome encodes stable modification information on top of the invariant genomic sequence (Fig. 1a)[13–15]. Likewise, a system that encodes data epigenetically might suffice for the long-term storage of information on the same DNA sequence (Fig. 1a). However, current in vitro parallel writing of epigenetic information is not versatile enough in data selectivity[16,17]. Therefore, a framework that programs arbitrary epigenetic information on universal DNA is desirable for the purpose of synthesis-free DNA data storage.

On the other hand, DNA self-assembly has been well studied to reach accurate molecular programmability at nanoscale. DNA sequence programming has enabled parallel and precise control of supramolecular structures at a large scale. Previous research showcased the ability to program multi-micron crisscross DNA structures and wireframe DNA polyhedra using thousands of kinds of DNA strands in one reaction[18,19]. Inspired by natural epigenomic inheritance and by synthetic DNA self-assembly, we have developed an unconventional DNA data writing framework that allows arbitrary epigenetic information bits (epi-bits) to be stably written in a parallel manner onto DNA templates based on DNA self-assembly guided enzymatic methylation (Fig. 1b–d).

System design and single-bit validation

Our framework proposes the strategy of DNA self-assembly guided enzymatic methylation to implement parallel and selective writing of epi-bits onto DNA templates with a premade set of DNA movable types and the methyltransferase DNMT1, much as typography to press information on papers[20] (Fig. 1b–d). First, a universal single-stranded

[1]School of Computer Science, Key Laboratory of High Confidence Software Technologies, Peking University, Beijing, China. [2]Center for Quantitative Biology, Peking University, Beijing, China. [3]School of Control and Computer Engineering, North China Electric Power University, Beijing, China. [4]2nd Physics Institute, University of Stuttgart, Stuttgart, Germany. [5]Max Planck Institute for Solid State Research, Stuttgart, Germany. [6]Center for Molecular Design and Biomimetics, Biodesign Institute, Arizona State University, Tempe, AZ, USA. [7]School of Molecular Sciences, Arizona State University, Tempe, AZ, USA. [8]These authors contributed equally: Cheng Zhang, Ranfeng Wu, Fajia Sun, Yisheng Lin. ✉e-mail: zhangcheng369@pku.edu.cn; qi@pku.edu.cn; long.qian@pku.edu.cn; hao.yan@asu.edu

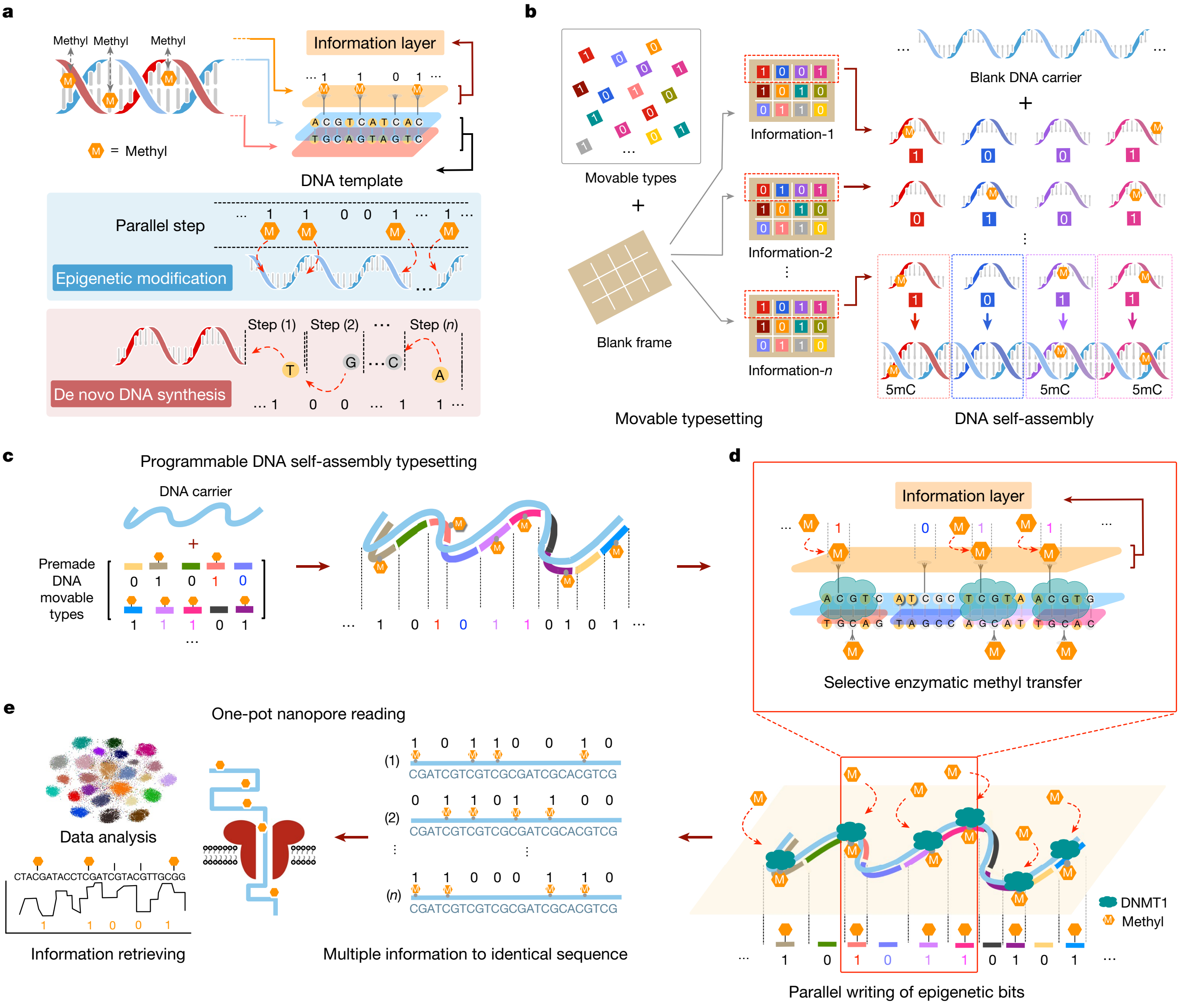


Fig. 1 | Schematics of the epi-bit DNA storage. a, Illustration of the mechanism of epigenetic information storage. **b**, Schematic for programming DNA movable types. **c**, Programmable DNA self-assembly typesetting. The modification information is typeset by programing DNA movable types carrying specific epi-bits. **d**, Parallel printing is performed by DNA self-assembly guided DNMT1 catalysis to selectively write epi-bits to the DNA templates. **e**, Nanopore sequencing of modified templates and collective methylation calling.

DNA (ssDNA) carrier and a library of complementary short ssDNA bricks are designed and prefabricated. Then, arbitrary epi-bit information is typeset by assembling the brick library onto the identical loading sequence of the DNA carriers (Fig. 1c). Next, base modifications (5-methylcytosine (5mC)) are stably 'printed' in a parallel manner on the DNA carriers through selective methylation by DNMT1 (Fig. 1d). Finally, numerous epi-bit streams stored on different carrier molecules are retrieved in high throughput by one-pot nanopore sequencing (Fig. 1e). The strategy enables the parallel writing of arbitrary data in DNA using premade nucleic acids instead of de novo synthesis. This enzymatic printing process potentially reduces the cost and time beyond the limit of chemical synthesis, and the accuracy of data writing is endowed by highly specific brick-template DNA assembly (Supplementary Fig. 1 and Supplementary Note 2).

To test in vitro epi-bit writing, we first implemented the selective transfer of a single epi-bit. In the design, a CpG dinucleotide as the methylation site is set in a ssDNA brick *p* or *q* with the cytosine methylated or unmethylated, respectively (Fig. 2a). DNMT1 specifically recognizes the hemimethylated site in DNA *p* and transfers a methyl group from *S*-adenosylmethionine (SAM) to the opposite cytosine in the DNA template, achieving the writing of epi-bit 1. By contrast, for the unmethylated brick *q*, no methyl group transfer happens and the DNA template remains in the epi-bit 0 state (Fig. 2a). To test the efficiency of selective epi-bit writing, we developed a molecular beacon where a template and a brick hybridized to bring a quencher and a fluorophore in proximity (fluorescent OFF state). Only when the template-brick complex contained a quadruple-methylated GCGC site, could it be cleaved by the restriction endonuclease GlaI, sending out the fluorescence signal[21]. Specifically, the single-methylated template b1 (G5mCGC) was tested against brick a1 (G5mCG5mC, epi-bit '1') or a2 (GCG5mC, epi-bit '0'), respectively (Fig. 2b). By gel and fluorescence assays, it is clear that only a1 induced methyl group transfer, indicating the precise selective writing of a single epi-bit (Fig. 2c,d and Supplementary Fig. 4).

For epi-bit reading, we explored nanopore sequencing (Fig. 2e). DNA template T1200 (1,200 nt) was designed with 101 CpG dinucleotides as methylation sites (Supplementary Figs. 6 and 7 and Extended Data Fig. 1), which were distributed into 24 nt segments each containing one to five CpG sites (Fig. 2f). After treatment with the highly efficient 5mC

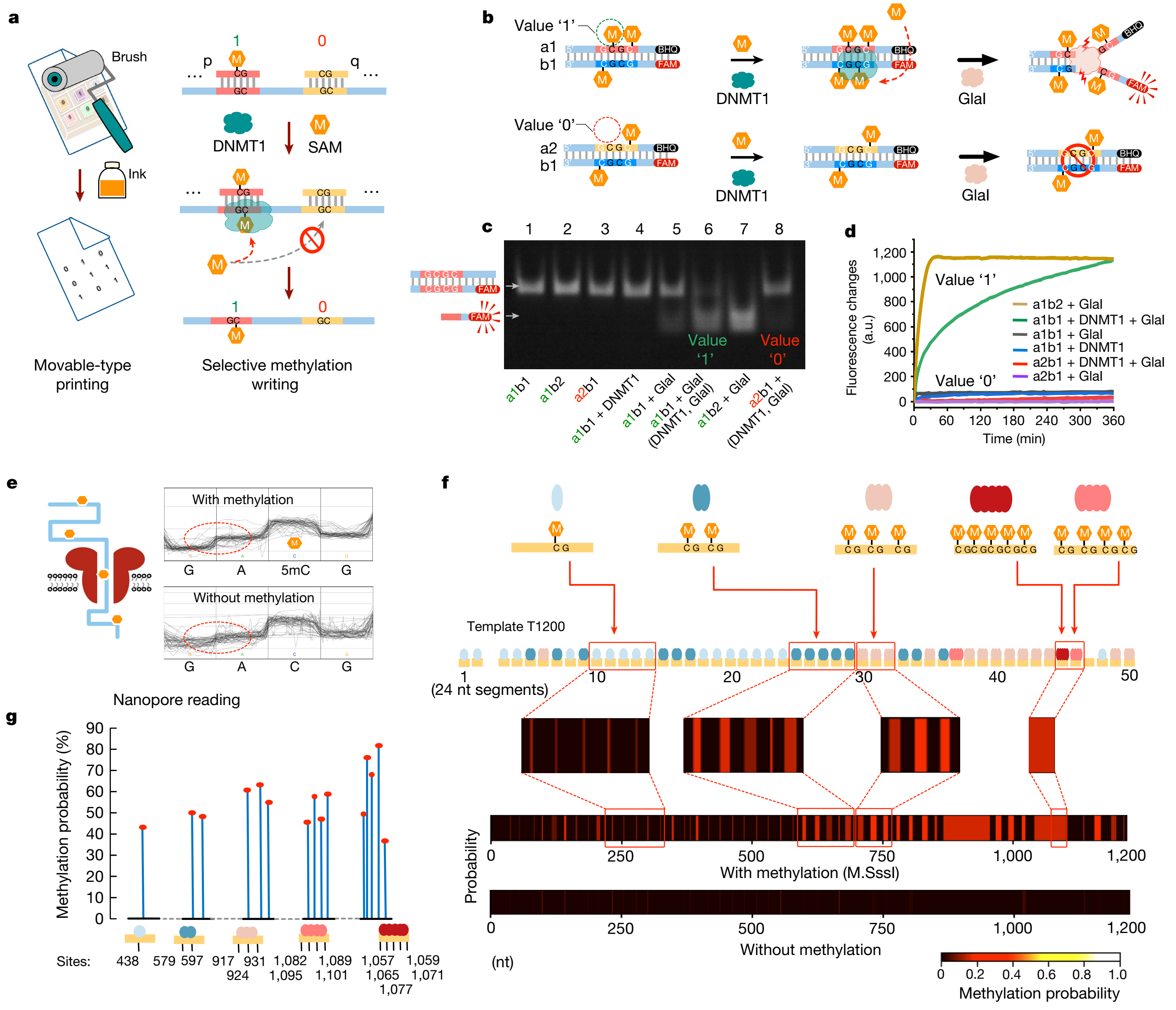


Fig. 2 | Design and verification of selective epi-bit writing. a, Schematic of the principle of selective methylation. **b**, Design of the single-bit methylation writing assay. **c**, Gel-shift assay detecting different methylation states. (DNA b2: positive control of double methylated DNA b1). **d**, Fluorescent detection of different methylation states. **e**, Illustration of epi-bit reading by nanopore sequencing. **f**,**g**, Methylation calling of the nanopore signals of templates treated by M.SssI (**f**) and DNMT1 (**g**) using Nanopolish and Megalodon, respectively. The true states of methylation are shown in cartoon representations. Methylation probabilities are shown as colour intensities (**f**) or bar heights (**g**).

de novo methyltransferase M.SssI[22], the nanopore current signals of fully-methylated T1200 exhibited significant signal shifts around the methylation sites (Fig. 2e). Methylation calling by Nanopolish reliably detected all segments even when the segment contained only one 5mC[23] (Fig. 2f). By contrast, almost no distinctive methylation was called from templates untreated with M.SssI. Selective epi-bit writing was also implemented with DNMT1 and 24 nt ssDNA bricks. The nanopore signals called by Megalodon[24] exhibited remarkable resolution in distinguishing closely spaced epi-bits within a 24 nt brick segment (Fig. 2g and Extended Data Fig. 2a–d).

Typesetting by DNA self-assembly

To program arbitrary epi-bit data, a DNA self-assembly-based typesetting strategy was developed (Fig. 3a(i)). A typical DNA movable type was designed as a 24 nt ssDNA brick with a CpG dinucleotide as the epi-bit site. The epi-bit value of movable types is governed by the presence (1) or absence (0) of the 5mC modification (Fig. 3a(ii)). By selecting specific combinations of DNA movable types to assemble with ssDNA carriers, the epi-bits on movable types were aligned in a parallel manner to establish a methylation pattern representing the coded information. Meanwhile, DNA carrier T960 (960 nt) was designed to store up to 36 bits of information in its loading sequence with a total of 72 kinds of movable type. We tested the operation by assembling T960 with 10 to 36 movable types. In gel results, DNA carriers shifted to low mobility bands stepwise with increasing numbers of movable types, indicating successful DNA typesetting (Fig. 3b and Extended Data Fig. 1).

As a demonstration, we typeset the ASCII code for letters 'DNA' (24 bits) to the T960 carrier. Subsequently, the sample was incubated with DNMT1 for selective printing, that is, the parallel transfer of methyl-coded information to the template strand. In fact, typesetting and enzymatic printing complete the cycle of data writing (Fig. 3a,c). After nanopore sequencing and methylation calling, the information was retrieved with negligible background noise (Fig. 3d,e and Extended

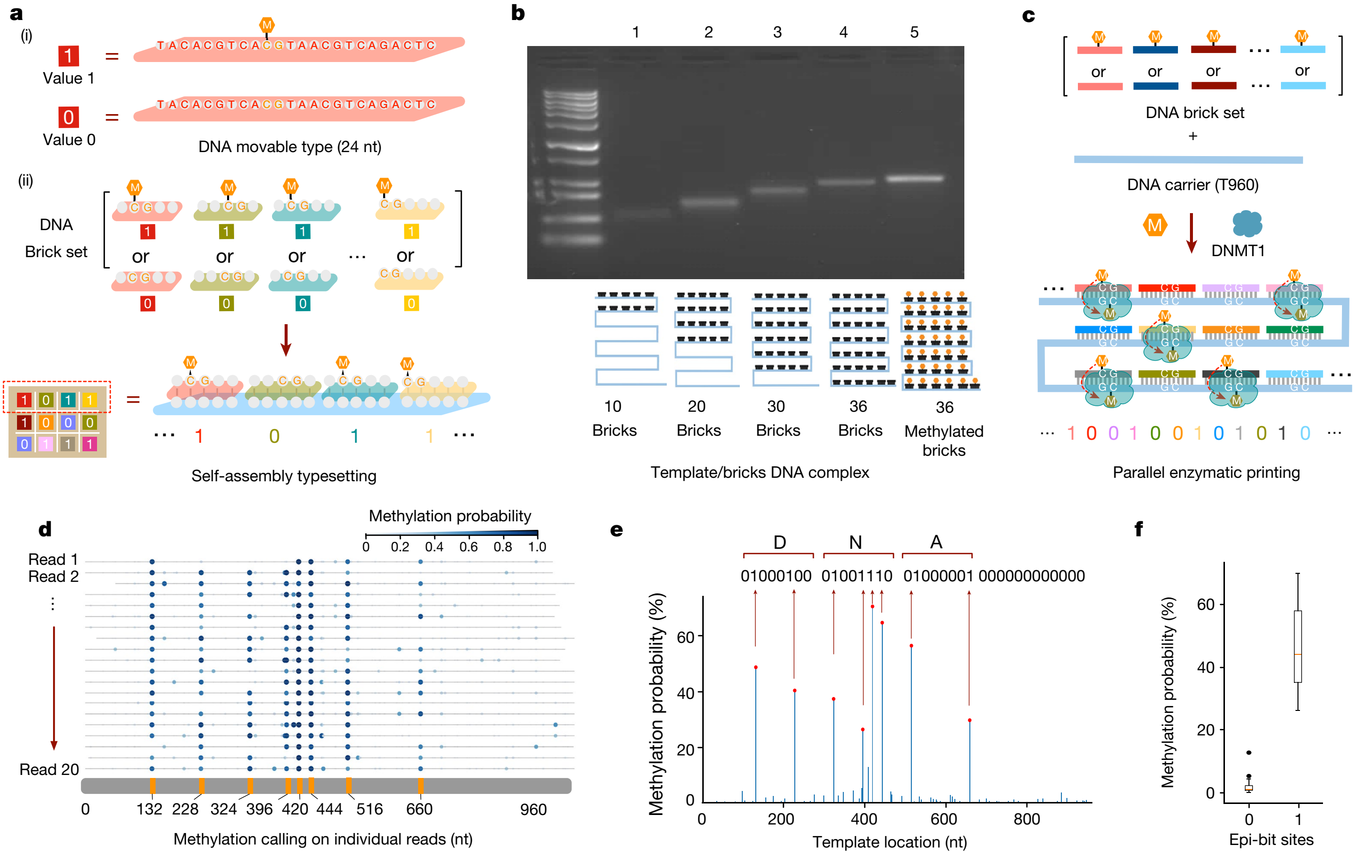


Fig. 3 | Programmable DNA typesetting and parallel epi-bit writing. **a**, Diagram of programmable self-assembly based typesetting. **b**, Gel results of the 960 nt DNA carriers assembled with varying numbers of DNA bricks. **c**, Diagram of parallel enzymatic epi-bit printing. **d**, Predicted methylation probabilities on 20 randomly selected sequencing reads of the DNA carrier encoding 'DNA'. **e**, Collective methylation probability retrieved at epi-bit sites encoding the information of 'DNA'. **f**, The predicted methylation probability distribution at epi-bit 0 versus epi-bit 1 sites ($n = 62$ for epi-bit 0, $n = 8$ for epi-bit 1).

Data Fig. 2a–d). As signal distributions were distinct for epi-bit 1 sites versus epi-bit 0 sites (Fig. 3f), it was possible to apply a threshold for accurate information retrieval. In addition, the transferred epi-bits endured heating at 65–95 °C, demonstrating the stability of epi-bits (Extended Data Fig. 2e,f and Supplementary Fig. 15).

Enlarged storage by sequence barcodes

Because nanopore sequencing resolves both sequence and modification information, we first scaled up storage by writing 800-bit information onto 25 barcoded DNA carriers with a universal template sequence using the same set of 72 movable types (Fig. 4a–c). Eight structure drawings of DNA bases and their modified derivatives were first transformed by bitmap coding (10 × 10) and sparsified to reduce the number of epi-bit 1s (Fig. 4a and Supplementary Fig. 16). The data were then allocated to different DNA carriers, which were designed to have three consecutive regions: (1) a 59 nt DNA extension to buffer for unstable nanopore signals; (2) a unique 20 nt barcode for indexing; and (3) a universal 960 nt template as the loading sequence for epi-bit writing (Fig. 4b). The barcode and template regions were generated independently, aligned by a ssDNA linker and ligated (Fig. 4c, Supplementary Fig. 18 and Extended Data Fig. 3). For writing, barcoded DNA carriers were treated by DNMT1 as guided by select DNA movable types representing the allocated data (Fig. 4c,d).

For data retrieval, one-pot nanopore sequencing of mixed DNA carriers yielded read signals that were first grouped by the methylation pattern, screened for the correct barcodes and then called for methylation collectively (Extended Data Fig. 4). Figure 4e,f shows the recovery of all original images from sequenced epi-bit streams. Comparison of the methylation probability distribution at each epi-bit site revealed differential discriminability between 0 and 1 (Extended Data Fig. 5), probably attributed to local context effects on methylation efficiency, variant self-assembly efficiencies owing to the DNA sequence design, or contextual signal detection in nanopore reads[25]. From these distributions, the methylation calling threshold was obtained as 0.345, which resulted in an error rate of 0.625% in one-pot sequencing (Fig. 4g).

Besides errors in methylation calling, errors could have originated from the promiscuous ligation by T4 ligase between barcodes and templates (Methods). To analyse relative contributions of the two error sources, k-means clustering ($k = 25$) of reads based purely on methylation patterns was re-analysed (Supplementary Fig. 28 and Supplementary Note 3). In Fig. 4h, each cluster corresponds to one segment of data written on a specific barcoded carrier. An average of 25.5% barcodes in the clusters were identified incorrectly (Supplementary Note 3). We simulated barcode mixing numerically due to methylation miscalling at the experimentally measured per-read miscalling rate $E_{\mathrm{meth}} = 0.3$ per site. After removing the effect of miscalling, the promiscuous ligation rate was estimated to be $E_{\mathrm{lig}} = 0.9\%$ (Fig. 4i). Nonetheless, without the reference of methylation patterns, information retrieval based solely on barcodes had an error rate of a mere 1.75%.

Data storage at high-bit parallelism

To leverage the parallel nature of the epi-bit writing mechanism, we implemented large-scale data storage with several DNA templates and denser epi-bit sites (Fig. 5a, Supplementary Fig. 39 and Extended Data Fig. 6). We define the bit parallelism, an essential feature of data

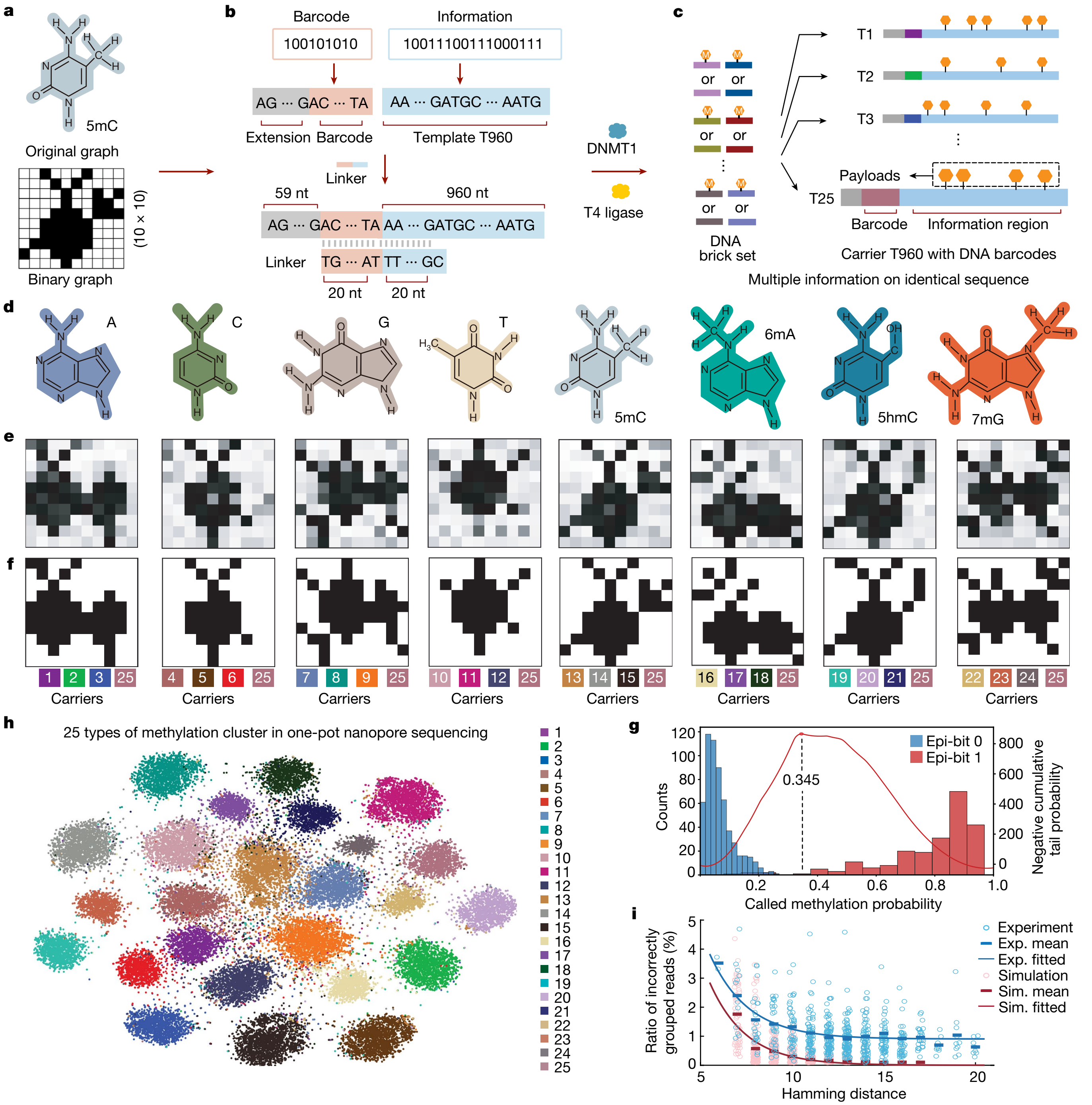


Fig. 4 | Enlarged epi-bit data storage and data retrieval by one-pot sequencing. a, Preprocessing of the original image data. **b**, Preparation of barcoded DNA carriers. **c**, Writing of epi-bit information onto several DNA carriers. **d**, Stored images of DNA bases and derivatives. 5hmC, 5-hydroxymethylcytosine; 6mA, N^6-methyldeoxyadenosine; 7mG, N^7-methylguanine. **e**, Images recovered by called methylation probabilities from nanopore sequencing reads (greyscale). **f**, Binarized image recovery based on the threshold of methylation calling. Coloured bars below each image indicate the indices of carriers used to encode the image. **g**, Determination of the threshold for methylation calling. The threshold was chosen to minimize the total overlap (potential miscalling) between dual-normal distributions fitted to methylation calling data at each epi-bit site. **h**, *k*-means clustering of all sequencing reads based on called epi-bit patterns. Patterns on each cluster are mapped to the original coded templates. **i**, Total barcode mixing fractions (representing read misalignment errors) as a function of the hamming distance between pairs of epi-bit patterns. The simulated (sim.) data (red circles) fitted to $y = \exp(-0.5x + 3.8)$. The experimental (exp.) data (blue circles) fitted to $y = \exp(-0.48x + 3.7) + 0.9$.

writing, as the number of bits written in a single minimal reaction per data-writing cycle. For example, de novo synthesis has a bit parallelism of around 1 (Supplementary Fig. 1, at the average density of 1.08 bits per base), whereas the enlarged storage experiment had a bit parallelism of 32. In the next set of large-scale experiments, five kinds of 1,300 nt DNA template (L1–L5) were designed to selectively hybridized with 175 DNA bricks each bearing two epi-bit (from a set of 700 movable types), achieving 350-bit parallelism per writing reaction (Fig. 5a(ii)).

With this design, we stored 269,337 bits including the image of a tiger rubbing from the Han dynasty in ancient China and the coloured picture of a panda (Fig. 5a(i) and Extended Data Fig. 6). Notably, reactions were labelled by epi-bit barcodes instead of ligated DNA barcodes to

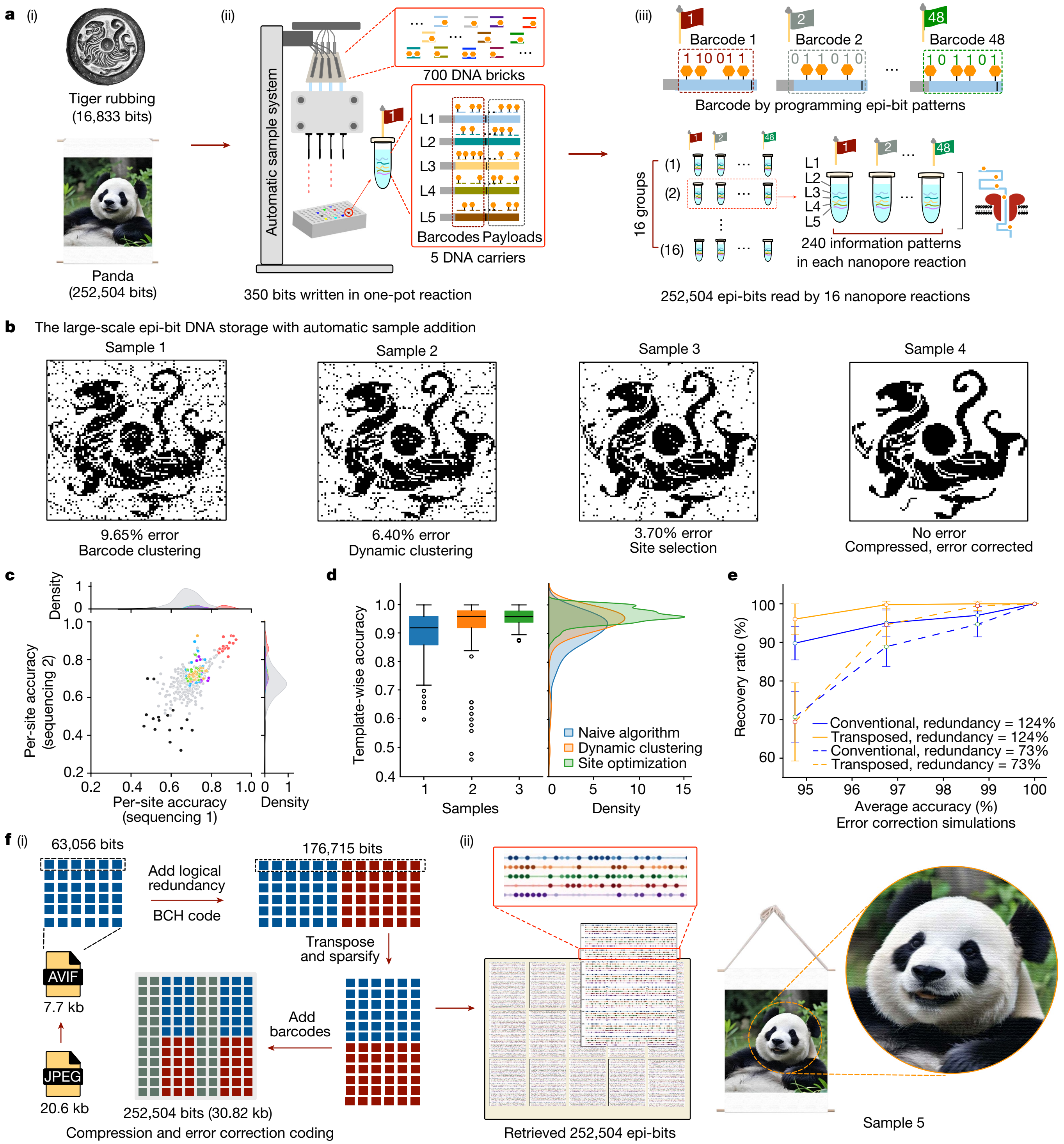
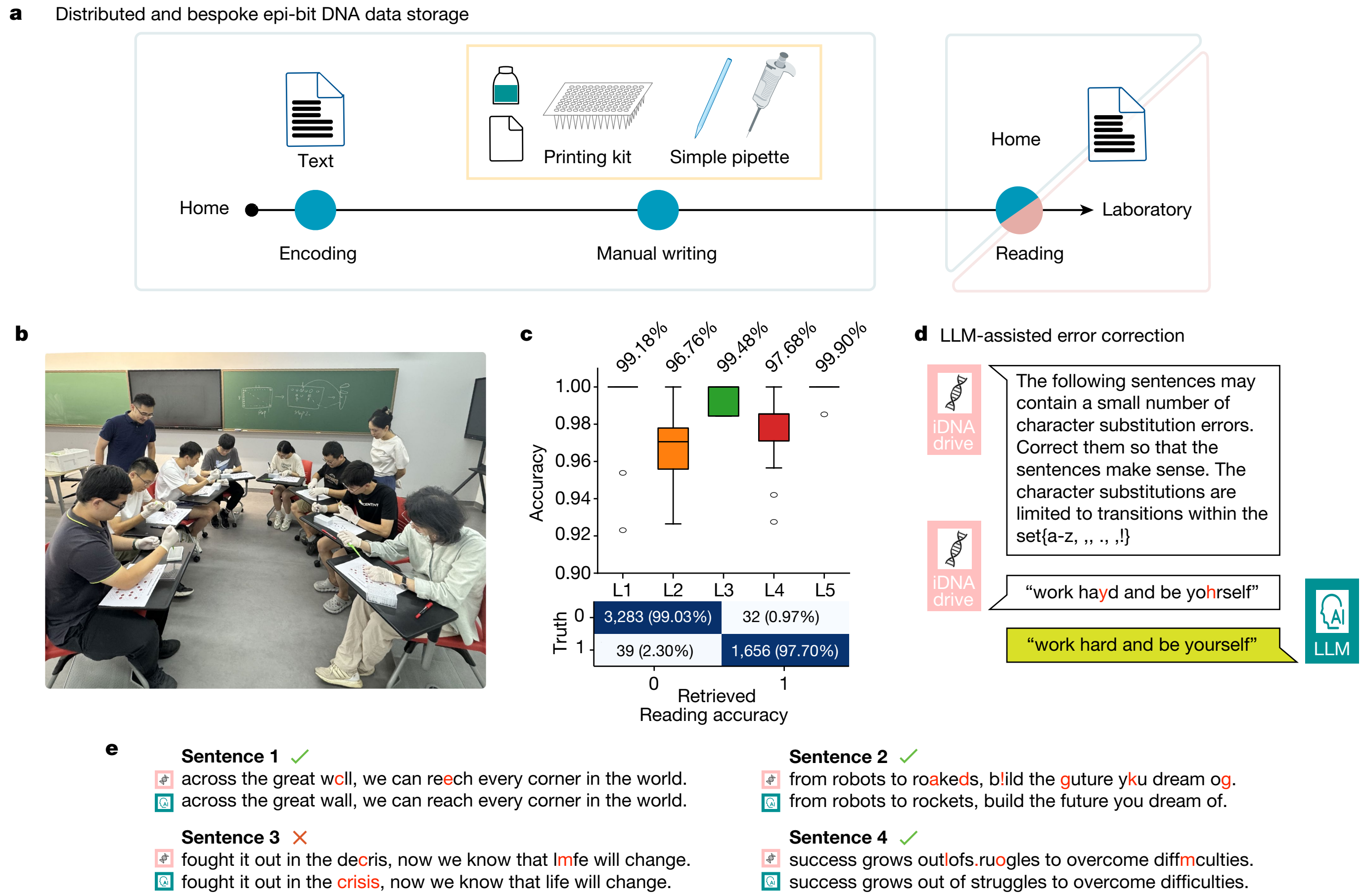


Fig. 5 | Large-scale storage with high-bit parallelism by epi-bit barcodes. **a**, Schematic pipeline of large-scale epi-bit DNA storage. **b**, Recovered tiger rubbing images of samples 1 to 4 with stepwise improved writing-reading pipelines. **c**, Per-site epi-bit accuracy between two independent sequencing samples of the originally encoded epi-bit data of the tiger rubbing. **d**, Distribution of template-wise accuracy of samples 1 to 3 ($n = 213$ for sample 1, $n = 213$ for sample 2, $n = 223$ for sample 3). **e**, Simulations of the error correction capabilities for different error correction configurations of the BCH code for the storage of the panda image. For each point, data are shown as mean ± s.e.m. for 100 independent simulations. **f**, Compression and error correction coding scheme for panda image (i), and a schematic of the retrieved epi-bits on sequencing reads along with the restored image (ii).

reduce complexity and streamline experimental protocols (Fig. 5a(iii) and Extended Data Fig. 6). An automatic liquid handling platform was used to typeset large-scale data at a speed of approximately 40 bits s^{-1} (Fig. 5a(ii)). With subsequent DNMT1 treatment, data were printed in parallel onto DNA templates (Fig. 5a(iii)).

We first stored the tiger rubbing image by bitmap encoding (16,833 bits distributed to 48 barcodes; Fig. 5b, sample 1). After nanopore reading, collective methylation calling on barcode-separated information (barcode clustering method) yielded an accuracy of 90.35%, due primarily to the noisy single-read methylation probabilities at barcode



The following sentences may contain a small number of character substitution errors. Correct them so that the sentences make sense. The character substitutions are limited to transitions within the set{a-z, ,, ., ,!}

"work hayd and be yohrself"

"work hard and be yourself"

Sentence 1 ✓
across the great wcll, we can reech every corner in the world.
across the great wall, we can reach every corner in the world.

Sentence 2 ✓
from robots to roakeds, b!ild the guture yku dream og.
from robots to rockets, build the future you dream of.

Sentence 3 ✗
fought it out in the decris, now we know that lmfe will change.
fought it out in the crisis, now we know that life will change.

Sentence 4 ✓
success grows outlofs.ruogles to overcome diffmculties.
success grows out of struggles to overcome difficulties.

Fig. 6 | Bespoke and distributed epi-bit DNA storage. a, Writing and reading pipeline of iDNAdrive. **b**, Photograph of student volunteers writing epi-bit data in the classroom. **c**, Epi-bit retrieval accuracy of each template (top, *n* = 15 for each template) and on epi-bit 0 and 1 sites (bottom). The analyses were on all sequenced samples corresponding to 15 stored texts. **d**, Correction of character substitution errors by prompting large language models for the recovery of stored texts. Red letters indicate character mistakes. **e**, Four examples of retrieved and recovered texts.

sites (Fig. 5b, sample 1, Supplementary Fig. 39 and Extended Data Fig. 7). Therefore, we sought to improve read classification by an iterative dynamic clustering algorithm that also consulted data at information sites, which increased the accuracy to 93.60% (Fig. 5b, sample 2 and Extended Data Fig. 7a). Next, as we observed significant but consistent variations in per-site methylation discriminability, site selection was implemented, in which sites with low accuracies (2.05% of all sites) were discarded from coding and the sites ranked top in accuracy in each template served as the barcode sites for a third storage experiment (Fig. 5c and Extended Data Fig. 7b,c). As a result, site selection improved data retrieval accuracy significantly to an average of 96.30%, with much narrower distributions among templates (Fig. 5b, sample 3 and Fig. 5d). Finally, we developed a compression and error correction coding significantly scheme to achieve full data recovery from one-pot nanopore sequencing reads (Fig. 5b, sample 4 and Supplementary Figs. 40 and 41). In silico evaluation of various error correction parameters suggested transposition effectively dispersed clustered errors to recover the original information from up to 5.3% errors (Fig. 5e).

Finally, we integrated the above strategies to store the panda image of 252,504 bits (15-fold of the tiger image) after compression, error corrective encoding and barcoding (Fig. 5f(i)). The data were written in 756 reaction wells by the automatic platform. For retrieval, every 48 wells (240 epi-bit patterns stored on DNA templates) were sequenced in one batch and underwent dynamic clustering and site-specific threshold selection, yielding an accuracy of 97.47% (Extended Data Fig. 8). With error corrective decoding, the image was restored perfectly (Fig. 5f(ii) and Supplementary Figs. 42–53).

Distributed and bespoke epi-bit storage

Owing to the low cost and high throughput of commercial DNA synthesis, many oligonucleotide-based DNA storage services adopt a centralized mode that requires the user to send the data to be stored to large-scale DNA synthesis platforms[26]. However, considering inefficient data communications and possible data breaches during electronic transmission, an alternative, distributed data storage mode has been under active research[27]. The simplistic and highly parallel nature of the epi-bit storage framework enables such a mode. It allows users to write their personal data into DNA by themselves, and the data remain undisclosed until the DNA are sequenced (Fig. 6a). To demonstrate this, we devised a distributed and bespoke DNA storage experiment named iDNAdrive.

In iDNAdrive, 60 student volunteers with diverse academic backgrounds were invited to manually store pieces of text of their choice (around 5,000 bits in total) in epi-bits using a simple epi-bit writing kit in a classroom (Fig. 6b and Supplementary Fig. 55). After manual typesetting and printing, the samples, barcoded for each volunteer group, were collected and sequenced. Analysis of the nanopore reads revealed an epi-bit error rate of 1.42% (Fig. 6c), owing to a subdued barcode misalignment rate (3.67%; Supplementary Note 3) and controllable manual sampling errors (Supplementary Fig. 58). This result underpinned the robustness of the data writing procedure. The sporadic substitution errors in the decoded texts were further rectified by large language model based text inference (Fig. 6d and Supplementary Fig. 59). Finally, 12 of the stored

15 texts were restored successfully and returned securely to the group that wrote them using an online server (Fig. 6e and Supplementary Fig. 60).

Discussion

Many elaborate studies have tackled oligonucleotide-based DNA storage, putting considerable efforts to increase the coding density and reduce the cost of de novo synthesis to make the writing cost commercially viable[1,8,28,29]. Enzymatic synthesis is explored as well for its potential to surpass phosphoramidite chemistry in speed and cost, but its selectivity and efficiency have remained key challenges[30,31]. The epi-bit storage framework adopts a fundamentally different mechanism, leveraging programmable DNA self-assembly and selective enzymatic methylation in synergy to enable parallel data writing on DNA. Compared with de novo synthesis-based storage, the epi-bit strategy possesses unique characteristics. First, it represents a parallel 'printing' process with prefabricated movable types and carriers. Therefore, information is written at the mere expanse of 'brush' and 'ink', that is, methyltransferase and SAM, respectively (Fig. 2a), which scales directly with the stored bits rather than the coding density (Supplementary Fig. 1 and Supplementary Note 2). In fact, at 10 µl reaction volumes and commercial enzyme prices, the cost of epi-bit writing is estimated to be within tenfold of that of de novo synthesis at high coding densities (Supplementary Fig. 1). Second, the parallelism of epi-bit writing exists both at the molecular level (the bit parallelism) and at the operational level (reaction throughput) (Supplementary Note 1). In the future, the fundamentally high-bit parallelism (approximately 324-fold of de novo synthesis) can be combined with high reaction throughput technologies (for example, massively parallel micro-inkjet printing) to achieve superior writing speeds[8,32] (Supplementary Note 1). These aspects attest the scalability of the epi-bit data storage framework. The storage of around 275,000 bits of data is a notable demonstration among the very few reported unconventional DNA storage strategies that reached more than 10 kB data sizes (Supplementary Table 1). Finally, the successful implementation of iDNAdrive underscores the practicality of the epi-bit framework, as well as the potential for developing desktop DNA printers meeting distributed and private data storage needs[27].

In our experiments, various error sources exist in the epi-bit storage workflow (that is, synthesis, storage and reading; Supplementary Table 2 and Supplementary Note 3). Nonetheless, the dynamic algorithm could resolve up to 240 different methylation information patterns at a high resolution within each sequencing batch (Supplementary Fig. 44). Whereas the single-read methylation calling error is around 30%, it decreased to 4–6% as the read coverage increased to over 10× (Supplementary Note 3). We achieved 97.5–99.4% sequencing retrieval accuracies among the experiments as governed by read clustering precision, which is prone to barcode mis-ligation errors (in the storage of base drawings) or errors in the epi-bit barcodes (in the storage of tiger rubbing and panda). In the future, more precise methylation detection algorithms and sequencing platforms may be integrated to boost the read throughput much beyond the current limit. To demonstrate a functionally complete data storage system, we also conducted preliminary experiments for copying and randomly accessing epi-bit data with Bst-mediated isothermal DNA replication (Supplementary Figs. 61–62). In this regard, alternative strategies such as base conversion[33] in conjunction with PCR and regular sequencing is worth exploring as well.

With DNA data storage entering the dawn of commercialization[26,34], the epi-bit framework demonstrates potential directions in parallel molecular information storage with prefabricated modularity. Extensive research routes are rolled out. For example, optimization of sequence designs and methyltransferase efficiencies[17,35] will enable robust and precise data writing. Storage density can be increased further by incorporating a variety of DNA modifications[36] matched with accurate detection methodology[37–39]. Finally, combining DNA self-assembly assisted programming and myriad enzymatic modifications, it is possible to realize diverse DNA storage and computation functions for practical and functionalized molecular data systems.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-024-08040-5.

1. Erlich, Y. & Zielinski, D. DNA fountain enables a robust and efficient storage architecture. *Science* **355**, 950–954 (2017).
2. Banal, J. L. et al. Random access DNA memory using Boolean search in an archival file storage system. *Nat. Mater.* **20**, 1272–1280 (2021).
3. Organick, L. et al. Random access in large-scale DNA data storage. *Nat. Biotechnol.* **36**, 242–248 (2018).
4. 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).
5. Service, R. F. DNA could store all of the world's data in one room. *Science* https://doi.org/10.1126/science.aal0852 (2017).
6. Jones, N. How to stop data centres from gobbling up the world's electricity. *Nature* **561**, 163–166 (2018).
7. Meiser, L. C. et al. Reading and writing digital data in DNA. *Nat. Protoc.* **15**, 86–101 (2020).
8. Nguyen, B. H. et al. Scaling DNA data storage with nanoscale electrode wells. *Sci. Adv.* **7**, eabi6714 (2021).
9. Service, R. F. New way to write DNA could turbocharge synthetic biology and data storage. *Science* https://doi.org/10.1126/science.aav6033 (2018).
10. Chen, K. et al. Digital data storage using DNA nanostructures and solid-state nanopores. *Nano Lett.* **19**, 1210–1215 (2019).
11. Chen, K., Zhu, J., Bošković, F. & Keyser, U. F. Nanopore-based DNA hard drives for rewritable and secure data storage. *Nano Lett.* **20**, 3754–3760 (2020).
12. Zhang, Y. et al. DNA origami cryptography for secure communication. *Nat. Commun.* **10**, 5469 (2019).
13. Raiber, E. A., Hardisty, R., Delft, P. V. & Balasubramanian, S. Mapping and elucidating the function of modified bases in DNA. *Nat. Rev. Chem.* **1**, 0069 (2017).
14. Spruijt, C. G. & Vermeulen, M. DNA methylation: old dog, new tricks? *Nat. Struct. Mol. Biol.* **21**, 949–954 (2014).
15. Jones, P. A. & Liang, G. Rethinking how DNA methylation patterns are maintained. *Nat. Rev. Genet.* **10**, 805–811 (2009).
16. Tabatabaei, S. K. et al. DNA punch cards for storing data on native DNA sequences via enzymatic nicking. *Nat. Commun.* **11**, 1742 (2020).
17. Mayer, C., McInroy, G. R., Murat, P., Delft, P. V. & Balasubramanian, S. An epigenetics-inspired DNA-based data storage system. *Angew. Chem. Int. Ed.* **55**, 11144–11148 (2016).
18. Zhou, Y., Dong, J. & Wang, Q. Fabricating higher-order functional DNA origami structures to reveal biological processes at multiple scales. *NPG Asia Mater.* **15**, 25 (2023).
19. Wintersinger, C. M. et al. Multi-micron crisscross structures grown from DNA-origami slats. *Nat. Nanotechnol.* **18**, 281–289 (2023).
20. Heber, J. Print and perish? *Nat. Mater.* **7**, 512–514 (2008).
21. Wood, R. J., McKelvie, J. C., Maynard-Smith, M. D. & Roach, P. L. A real-time assay for CpG-specific cytosine-C5 methyltransferase activity. *Nucleic Acids Res.* **38**, e107 (2010).
22. Liu, Q. et al. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nat. Commun.* **10**, 2449 (2019).
23. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
24. Oxford Nanopore Technologies. Megalodon. GitHub https://github.com/nanoporetech/megalodon (2022).
25. Adam, S. et al. DNA sequence-dependent activity and base flipping mechanisms of DNMT1 regulate genome-wide DNA methylation. *Nat. Commun.* **11**, 3723 (2020).
26. Roth, E. This $1,000 card can store a message in DNA. *The Verge* www.theverge.com/2023/12/4/23987748/biomemory-card-store-message-dna (2023).
27. IARPA funds team involving DNA Script, the Broad Institute, and Harvard University for DNA data storage research and development. *DNA Script* https://go.nature.com/3XRkdfI (2020).
28. 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).
29. Ceze, L., Nivala, J. & Strauss, K. Molecular digital data storage using DNA. *Nat. Rev. Genet.* **20**, 456–466 (2019).
30. Fan, C., Deng, Q. & Zhu, T. Bioorthogonal information storage in L-DNA with a high-fidelity mirror-image Pfu DNA polymerase. *Nat. Biotechnol.* **39**, 1548–1555 (2021).
31. Lee, H. et al. Photon-directed multiplexed enzymatic DNA synthesis for molecular digital data storage. *Nat. Commun.* **11**, 5246 (2020).
32. Roquet, N. et al. DNA-based data storage via combinatorial assembly. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.20.440194 (2021).

33. Arslan, S. et al. Sequencing by avidity enables high accuracy with low reagent consumption. *Nat. Biotechnol.* **42**, 132–138 (2024).
34. Shankland, S. Startup packs all 16GB of Wikipedia onto DNA strands to demonstrate new storage tech. *CNET* https://go.nature.com/3XEd54O (2019).
35. Liu, C. et al. DNA 5-Methylcytosine-specific amplification and sequencing. *J. Am. Chem. Soc.* **142**, 4539–4543 (2020).
36. Tabatabaei, S. K. et al. Expanding the molecular alphabet of DNA-based data storage systems with neural network nanopore readout processing. *Nano Lett.* **22**, 1905–1914 (2022).
37. Tourancheau, A., Mead, E. A., Zhang, X. & Fang, G. Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nat. Methods* **18**, 491–498 (2021).
38. Wang, Y. et al. Identification of nucleoside monophosphates and their epigenetic modifications using an engineered nanopore. *Nat. Nanotechnol.* **17**, 976–983 (2022).
39. Bendixen, K. K. et al. A qPCR technology for direct quantification of methylation in untreated DNA. *Nat. Commun.* **14**, 5153 (2023).

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

Methods

Design of DNA bricks and templates

The DNA templates (T960, T1200, L1, L2, L3, L4 and L5) and bricks were designed by a computer-assisted method. First, the reported software[40] was used to generate a series of DNA brick sequences. By using the design software, a set of short DNA bricks with a length of 24 nt was generated without a four-repeat base domain, while keeping the hamming distance at least four and the content of CG between 40% and 60%. Second, we analysed the DNA bricks and templates manually to test the DNA sequences to avoid complex secondary structures and mismatches between bricks and templates (Supplementary Fig. 6).

Fluorescence assay

All DNA fluorescent experiments were performed in 1× Tris-acetate-EDTA (TAE)/Mg^{2+} buffer at 25 °C using real-time fluorescence PCR (CFX Connect Real-Time System, Bio-Rad). DNA b1 (or b2) was labelled with fluorophore (FAM) at 5′ ends and a1 (or a2) was labelled with quencher (BHQ-1) at 3′ ends. The FAM fluorescence signal was detected at 492 nm excitation and 518 nm emission. In a typical reaction, 30 µl solution was used for the detection. The time dependence of the fluorescence signals was normalized to make the initial value starts at zero. The detection time interval was 2–5 min. The fluorescence results were obtained by averaging the values from three replicates of the experimental results.

Polyacrylamide gel electrophoresis experiments

The reactions of the methyl transfer were verified using native polyacrylamide gel electrophoresis (PAGE); 12% gels were prepared with 1× TAE/Mg^{2+} buffer with 12.5 mM $MgCl_2$. All samples were run at 100 V for 1.5–2 h at 4 °C. After staining the polyacrylamide gels with Stain-All, the gels were imaged by a scanner (Canon LIDE 100). To detect FAM-modified DNA complexes, the gels were imaged directly under ultraviolet light without staining (Gel Image System Tanon-1600 or camera).

Agarose gel experiments

Asymmetric PCR[41] experiments were verified by agarose electrophoresis; 1% agarose was prepared with 1× TAE buffer that had been supplemented with 2.5 mM $MgCl_2$. All samples were run at 60 V for 1.5–2 h at room temperature. After staining the polyacrylamide gels with GelRed, the gels were imaged by Gel Image System Tanon-1600.

Preparation of ssDNA carriers

The designed ssDNA template carriers (T960 and T1200) were synthesized chemically and inserted into plasmid pUC57 (ref. 42). ssDNA templates were then prepared using an asymmetric PCR strategy, where the concentration ratio of forward to reverse primers is 60:1. The reactions were carried out in a PCR thermal cycler using the following protocol: 94 °C for 5 min, then 30 cycles of 94 °C for 30 s, 60 °C for 30 s and 72 °C for 1 min, thereafter 75 °C for 10 min, and finally kept at 25 °C. To verify whether the purified gel band is the authentic target ssDNA carrier, we used FAM-modified complementary DNA probes to hybridize the target ssDNA. After adding fluorescent probes for 2 h at room temperature, target DNA complexes were separated by agarose electrophoresis (60 V for 1.5–2 h). After fluorophore labelling, target gel bands were collected by cutting the agarose gel (Extended Data Fig. 1).

Programmable typesetting and methylation writing

The procedures of the standard methyl modification experiment were as follows: (1) a set of methylated and unmodified DNA brick strands was selected to mix with DNA carriers at 40:1 at a concentration of 0.4 µM in 1× TAE/Mg^{2+} buffer. Then, the sample was annealed in a PCR thermal cycler using the following protocol: 95 °C for 5 min, 65 °C for 30 min, 50 °C for 30 min, 37 °C for 30 min, 25 °C for 30 min and finally kept at 25 °C. (2) Methyltransferase DNMT1 (refs. 43,44), SAM, methyltransferase reaction buffer and 50% glycerol were added to the reaction solution. The samples were then incubated at 37 °C for 3 h, 65 °C for 20 min and finally kept at 25 °C. (3) The DNA complexes were gel purified for the next experiments.

GlaI-digestion-assisted detection of basic methylation writing

DNA complexes were produced by mixing DNA strands with equal molar concentrations (4 µM) in a 1× TAE/Mg^{2+} buffer. The annealing protocol was used for DNA complex assembly in fluorescence assay experiments. In fluorescence assay experiments, the reagents for basic methylation writing were mixed as follows: DNA complexes 1.5 µl, DNMT1 buffer 3 µl, SAM 3 µl, 50% glycerinum 3 µl, DNMT1 5 µl and H_2O 14.5 µl. In PAGE experiments, the 30 µl basic methylation writing reaction included: DNA complexes 6 µl, DNMT1 buffer 3 µl, SAM 3 µl, 50% glycerinum 3 µl, DNMT1 5 µl and H_2O 10 µl. After the basic methylation writing reactions, the mixture was incubated at 37 °C for 3 h, 3 µl GlaI buffer and the corresponding concentration of GlaI enzyme (5 U μl^{-1}) were added[21,45] (the volume of water was changed to keep the total volume constant at 30 µl). For fluorescent detection, the samples were placed in a fluorescence detector and reacted at 30 °C for 3 h. Fluorescence was detected every 3 min. For PAGE analysis, 12% PAGE was used and the voltage was kept at 90 V for 70 min.

High-throughput epi-bit DNA storage

The experimental procedure consisted of three parts: carrier T960 preparation, barcode connection and methyl writing. (1) Carrier preparation process: first, ssDNA carrier T960 was amplified from the plasmid by asymmetric PCR. Then the 5′ end of the carrier was phosphorylated by T4 polynucleotide kinase (PNK), followed by labelling of the target carrier T960 by fluorescent probes and purifying the carrier from agarose gels using the reagent kit (NucleoSpin Gel and PCR Clean-up, Mini kit). (2) Barcode connection: T960 carriers were specifically ligated with the 25 barcodes in individual tubes by adding T4 ligase at 25 °C for 30 min, followed by inactivation at 65 °C for 20 min. Then, 25 kinds of barcoded DNA carriers were purified by electrophoresis through a 1% agarose gel and recovery by the reagent kit. (3) Methyl writing: each of the 25 kinds of barcoded DNA carrier was mixed with a set of DNA bricks at a certain ratio and annealed. The methyl writing reaction was then carried out by adding DNMT1 methyltransferase at 37 °C for 3 h. Subsequently, the 25 kinds of carrier were mixed into one tube and purified using the reagent kit. Finally, the mixed sample of 25 kinds of carrier was prepared for nanopore sequencing.

Nanopore sequencing preparation

DNA samples were first prepared following the protocols of the Ligation Sequencing Kit (Oxford Nanopore Technologies (ONT), SQK-LSK109/SQK-LSK110) to construct the sequencing library, and then sequenced on MinION single-molecule sequencing device (ONT) by loading 50 fmol sample into a R9.4.1 flowcell. The device was operated using the bundled software MinKNOW to monitor running status. Base calling was done individually after sequencing using Guppy (ONT).

Methylation calling

Software Megalodon[23] was used for methylation calling. Megalodon is an ONT-developed analysis tool that is capable of calling modified bases with high precision by anchoring the information-rich base-calling neural network output to a reference DNA sequence. Megalodon predicts methylation at both the per-read and per-site level (by aggregating per-read results) based on the log probability of whether or not the base is modified. The primary Megalodon run mode requires Guppy base caller (v.4.0 and above), and appropriate Rerio model is recommended for the accurate modified base calls[46]. In experiments, we used Megalodon v.2.5.0 with Guppy v.5.0.16 with a 5mC calling model (res_dna_r941_min_modbases_5mC_CpG_v001) from Rerio, and chose the default probability cutoff (0.8) to predict DNA methylation.

Coding strategy to store images of modified nucleotide structures

Chemical structure drawings of four DNA bases and their modified derivatives were first converted into 10 × 10 bitmap pictures, and then flattened to obtain binary sequences with a length of 100 bits. Next, a random seed was used to generate a sequence of binary numbers of equal length, which was then used to perform bitwise XOR operation with the sequence of information, to obtain a new sequence with changed distribution of 1 s. This process was performed iteratively until a sufficiently sparse sequence was obtained, namely, the proportion of epi-bit 1 s in the sequence was less than one-third and the maximum number of consecutive epi-bit 1 s is 3. Each of the sparse sequences was finally written into a group of 960 nt DNA carriers with the same loading sequence (Supplementary Fig. 16).

Information retrieval from high-throughput epi-bit storage

The sequence representing the original digital information was a stream of binary number, and bit-by-bit XOR was used to sparsify the sequence. Therefore, each methylation site corresponded to one pixel in the original binary images. This correspondence produced a grey-scale image as the methylation predictions obtained from Megalodon are probabilities (0–100%). Restoring these images required de-sparsification. Specifically, for each bit (pixel), if 0 is used for performing XOR for an epi-bit site, the pixel value in grey-scale image is exactly the methylation probability. In the opposite circumstances, the value in grey-scale image is 1 minus the methylation probability of the corresponding methylation site (Supplementary Fig. 29). After these conversions, grey-scale images were binarized to bitmaps by setting an appropriate threshold.

Determination of the threshold of epi-bit calling

To determine the optimal threshold for calling epi-bits 1 and 0 from methylation probability data, we conducted analysis for the 32 methylation sites on the T960 DNA carriers individually. Specifically, for each methylation site, we assumed that both the probability values detected in nanopore sequencing of the status 0 (no methylation) and the status 1 (methylation) follow the Gaussian distribution. Therefore, the probability distribution detected from a methylation site was supposed to be the sum of two independent Gaussian distributions,

$$\mathrm{BG}_i(x,\alpha,\mu_{i1},\mu_{i2},\sigma_{i1},\sigma_{i2})=\alpha_i\frac{1}{\sigma_{i1}\sqrt{2\pi}}e^{-\frac{(x-\mu_{i1})^2}{2\sigma_{i1}^{\ 2}}}+(1-\alpha_i)\frac{1}{\sigma_{i2}\sqrt{2\pi}}e^{-\frac{(x-\mu_{i2})^2}{2\sigma_{i2}^{\ 2}}}$$

where for each methylation site i, μ_{i1} and μ_{i2} are the mean and s.d. of the two Gaussian distributions for status 0 and status 1, respectively, and σ_{i1} and σ_{i2} are the s.d. of the two Gaussian distributions for status 0 and status 1, respectively. α_i is the fraction of status 0 in all epi-bits printed at site i. To obtain the values of the above parameters, we fitted the above functions with the probability distribution obtained from each methylation site. It is worth noting that μ_{i1} should be close to 0 (as the unmethylated position is unlikely to generate a higher methylation signal), and μ_{i2} should be a higher value, otherwise the fitting will generate only a single peak. After fitting, each of the 32 methylation sites generated the corresponding double peak distribution (Extended Data Fig. 5). Next, we defined an objective function on threshold,

$$\mathrm{OBJ}(\mathrm{th})=-\sum_{i=1}^{32}\left(\int_0^{\mathrm{th}}\frac{1}{\sigma_{i1}\sqrt{2\pi}}e^{-\frac{(x-\mu_{i1})^2}{2\sigma_{i1}^{\ 2}}}+\int_{\mathrm{th}}^{1}\frac{1}{\sigma_{i2}\sqrt{2\pi}}e^{-\frac{(x-\mu_{i2})^2}{2\sigma_{i2}^{\ 2}}}\right)$$

Then, we maximize the objective function to obtain the optimal threshold.

Automatic sampling for data writing

Data writing of the pictures of the tiger rubbing and the panda were performed on an automated four-channel liquid handling system by HCSCI China. Briefly, the stock solution of 175 × 4 = 700 bricks was loaded into eight 96-well plates (source plates). Specific combinations of stock solutions were added to destination wells on 384-well plates, where each well held all bricks necessary to guide data writing on five templates L1–L5. A sampling sheet was prepared based on the epi-bit information of the data to be stored. The sampling sheet dictated the sampling scheme from the source wells to the destination wells. The liquid handling system then dispenses 500 nl of each source solution to each destination well according to the sampling sheet. Each destination well held 175 × 0.5 = 87.5 µl of mixture brick solution after the sampling was completed. The carriers, enzymes and reaction buffers were then added to all destination wells for data writing.

Coding strategies for data compression and error correction for the tiger rubbing and panda

The original image was first read as a binary stream. This binary stream was compressed. Next, BCH code was used to add logic redundancy to the compressed information. Specifically, the information was divided into groups. Each of these groups was used as information symbols to generate redundancy, which resulted in a coding matrix. Next, this matrix was transposed and flattened, resulted in a binary stream.

All the barcodes used for storing information were 20 bits in length. A seed barcode was first generated (for example, 01110101001011001001), and then barcodes with random bits were generated. Next, each barcode was verified to conform to the following rule. The valid barcode was recorded only when the minimum Hamming distance between it and all the recorded barcodes is greater than four. A total of 370 valid barcodes were selected, of which 250 barcodes possessed a 1 ratio of 40–60%. After site optimization (Extended Data Fig. 7), 16 sites were dropped and 5 × 20 = 100 sites were selected as barcode sites; thus, there were 234 sites for storing the image data. The compressed binary stream was divided into groups (234 bits per group). Depending on the barcode generation strategy, barcodes were selected randomly and assigned to groups. Finally, all groups were stored in wells.

The retrieved binary stream was first truncated, and then rearranged to a matrix. Next, this matrix was transposed, and each row of this matrix was used as a decoding unit. After BCH decoding, this matrix was flattened, resulted in a binary stream. Finally, the binary stream was visualized as the stored image.

Simulations of error correction capabilities in large-scale epi-bit DNA storage

In silico simulation was performed to test the capacity of error correction for epi-bit DNA storage, where the epi-bit information of tiger rubbing (Fig. 5e) and panda (Supplementary Fig. 42) were simulated independently. For all simulation, assuming that 50 bits of information was loaded on each DNA carrier, and errors were distributed independently on different DNA carriers. The error frequency was sampled from a pre-experiment, in which 48 wells (240 DNA carriers) were sequenced collectively with nanopore sequencing. The fluctuations of error rate were realized by adding or deleting single epi-bit error manually on DNA carriers.

Data availability

The data reported in this paper are available in the main text or the Supplementary Information. Source data are provided at https://bdainformatics.org/dataRepository, under Item 05. All other data are available on request.

Code availability

Code reported in this paper is available at https://bdainformatics.org/dataRepository, under Item 05. All other codes are available from the corresponding authors on reasonable request.

40. Zadeh, J. N. et al. NUPACK: analysis and design of nucleic acid systems. *J. Comput. Chem.* **32**, 170–173 (2011).
41. Veneziano, R. et al. Designer nanoscale DNA assemblies programmed from the top down. *Science* **352**, 1534 (2016).
42. Xie, X. et al. Engineering SARS-CoV-2 using a reverse genetic system. *Nat. Protoc.* **16**, 1761–1784 (2021).
43. Song, J., Teplova, M., Murakami, S. I. & Patel, D. J. Structure-based mechanistic insights into DNMT1-mediated maintenance DNA methylation. *Science* **335**, 709–712 (2012).
44. Song, J., Rechkoblit, O., Bestor, T. H. & Patel, D. J. Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. *Science* **331**, 1036–1040 (2010).
45. Sun, Y. et al. A novel restriction endonuclease GlaI for rapid and highly sensitive detection of DNA methylation coupled with isothermal exponential amplification reaction. *Chem. Sci.* **9**, 1344–1351 (2018).
46. Oxford Nanopore Technologies. Rerio. GitHub https://github.com/nanoporetech/rerio (2022).

Acknowledgements We would like to thank Y. Zhang, Z. Song and Y. Chen for their support during the course of this study. This work was supported by the National Key Research and Development Program of China (2023YFF1205600, 2021YFF1200103 to C.Z. and 2020YFA0906900 to L.Q.), the National Natural Science Foundation of China (62273008 to C.Z., 12090054 to L.Q. and 62073133 to Y. Liang), Pre-Research Project (31511090301 to C.Z.), the Fund of Kunpeng and Ascend Center of Excellence (Peking University), the Max Planck Fellow programme to N.L. and financial support from Arizona State University and the Alexander von Humboldt foundation (Humboldt Research Award) to H.Y.

Author contributions C.Z. and L.Q. conceived the research. R.W. and Y. Liang performed most of the experiments. F.S. and Y. Lin wrote programs and analysed data. F.S., Y. Liang and J.T. performed nanopore sequencing experiments. C.Z., L.Q., R.W., F.S. and Y. Lin participated in data analysis and discussions. C.Z., L.Q., N.L., H.Y. and Q.O. provided supervision. C.Z. and L.Q. wrote the original manuscript. C.Z., L.Q., N.L., H.Y., Q.O., R.W., F.S. and Y. Lin reviewed and edited the manuscript.

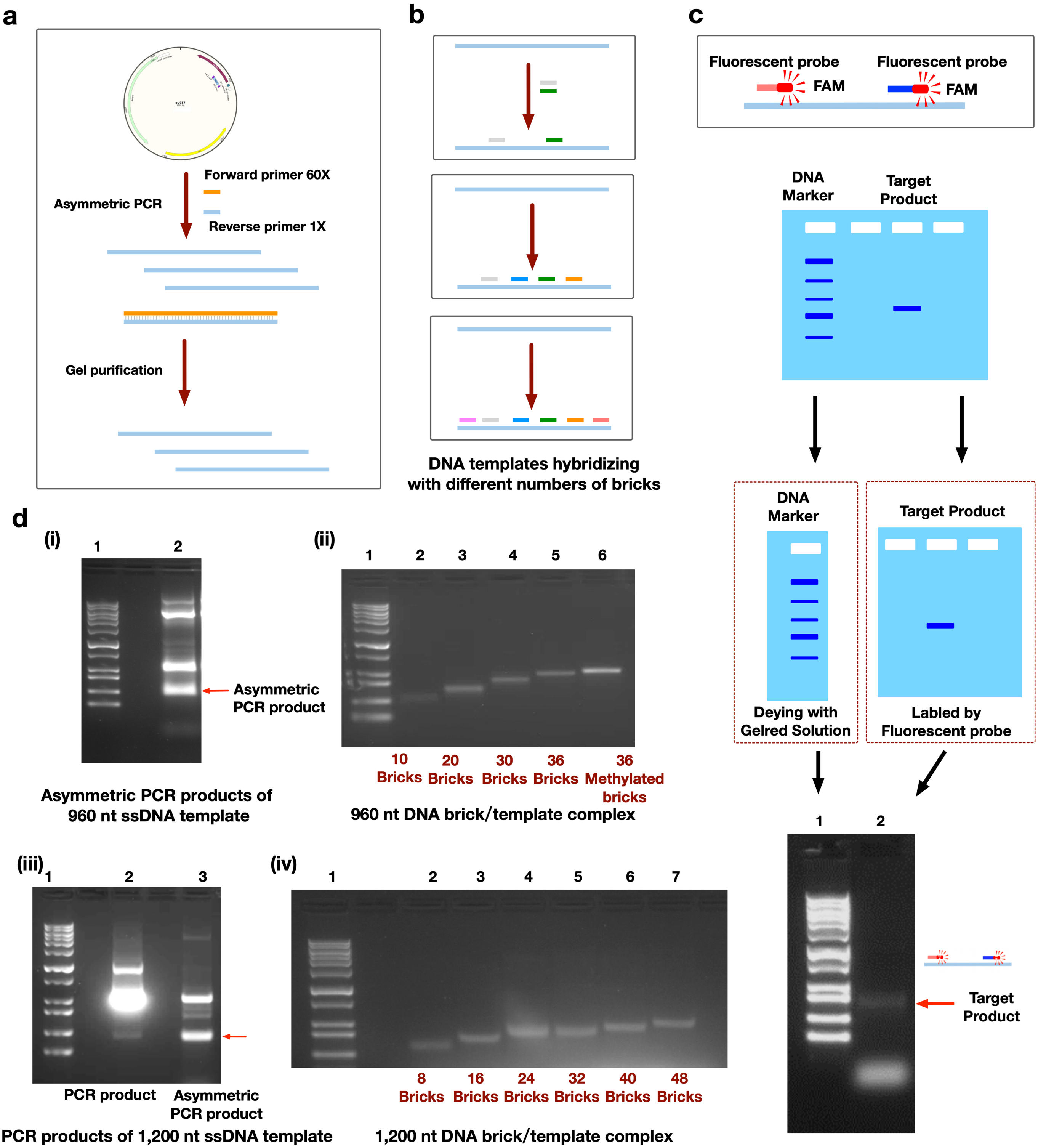
Competing interests The authors declare no competing interests.

Additional information

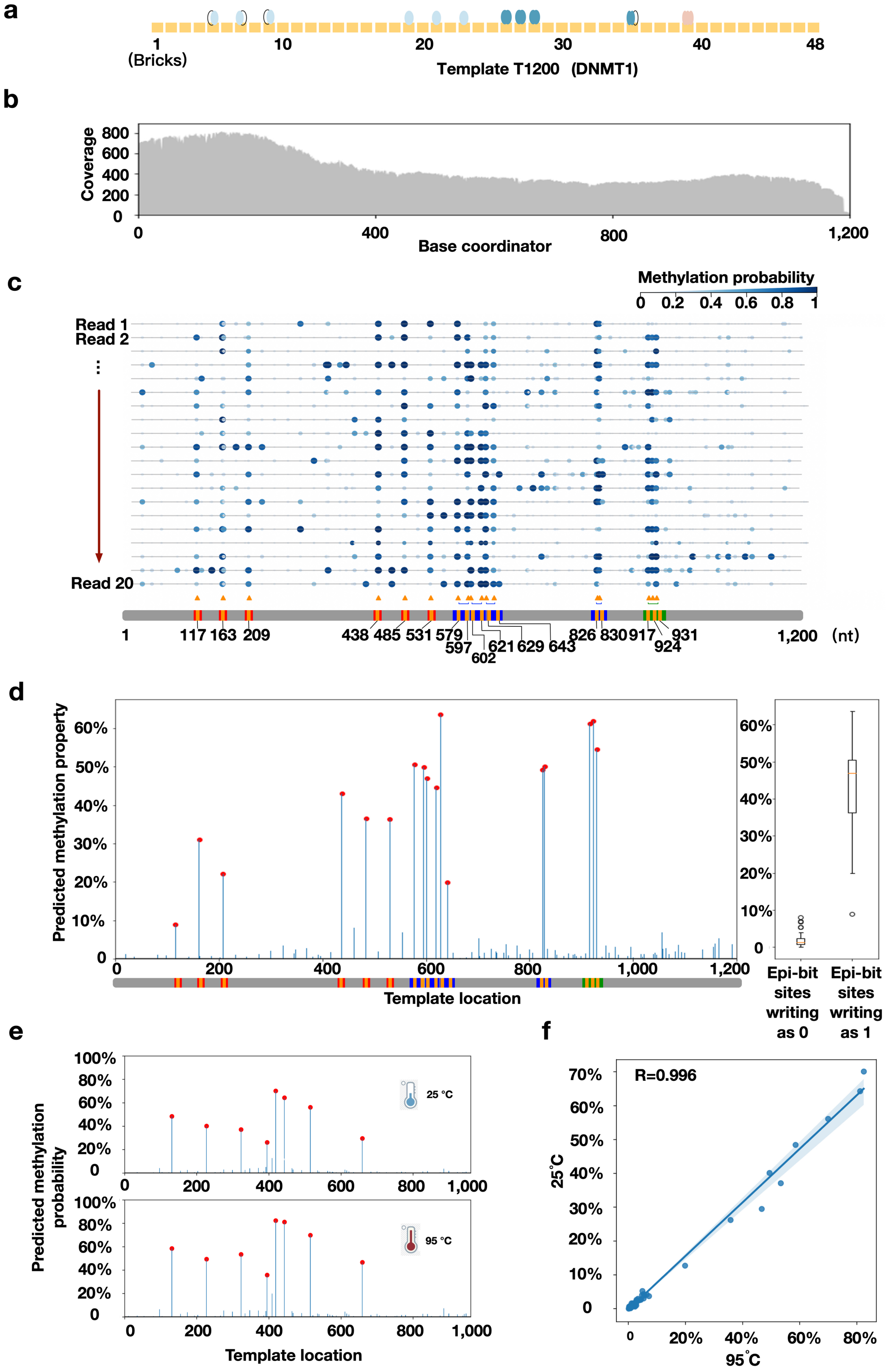
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-08040-5.

Correspondence and requests for materials should be addressed to Cheng Zhang, Qi Ouyang, Long Qian or Hao Yan.

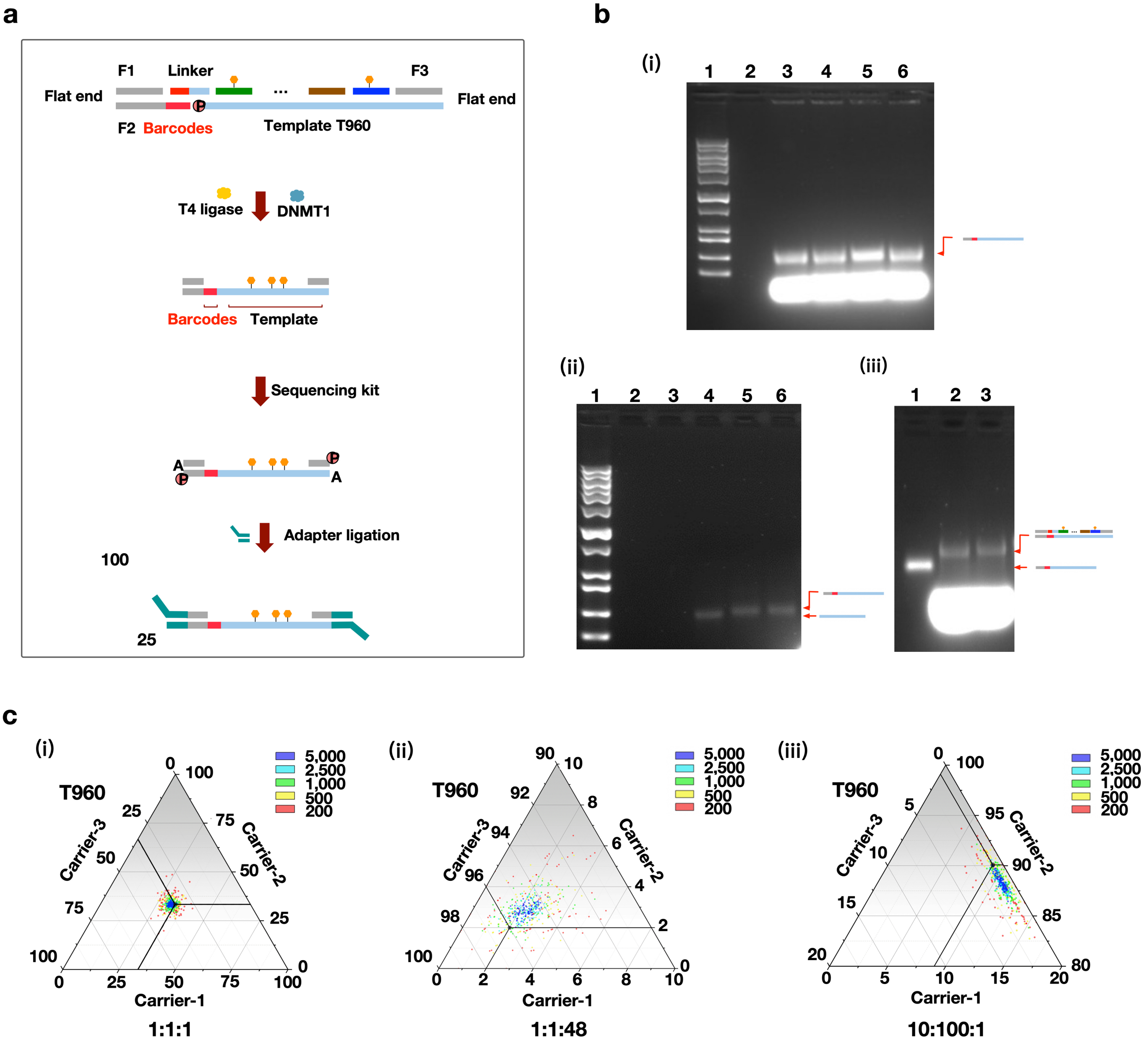
Peer review information *Nature* thanks Kun Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



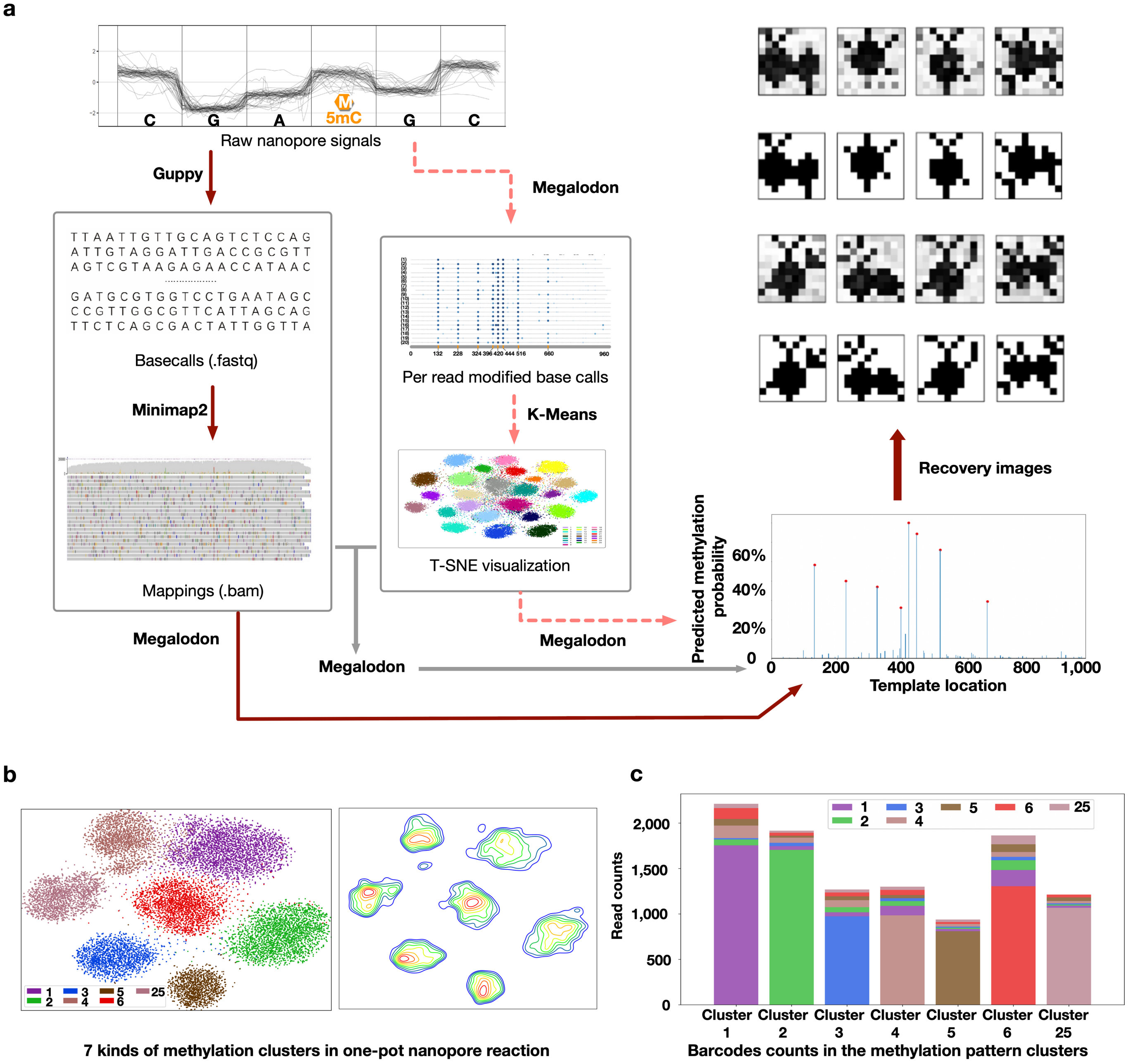
Extended Data Fig. 1 | The illustrations and results of the preparation of ssDNA carriers. a, The illustrations of the procedures to prepare ssDNA carriers. **b**, The illustrations of the DNA bricks-carrier complexes. **c**, The purification of the ssDNA carrier. The fluorophore modified DNA probes were used to hybridize the ssDNA carrier targets. To avoid using GelRed to dye the target DNA, the DNA ladders were individually cut as one agarose strip and subsequently soaked in GelRed solution. Then, the agarose strips of the ladders and ssDNA carriers were put together for detection under UV irradiation. The target ssDNA product was clearly observed in lane 2, and was then purified by using PCR clean up and gel extraction kit for the following methylation writing step. **d**, The gel results for the verification of the ssDNA carriers. (i) The gel results of the asymmetric PCR product of ssDNA carrier T960. (ii) The gel results of carrier T960 hybridized with 10, 20, 30 and 36 DNA bricks. (iii) The gel results of the asymmetric PCR product of ssDNA carrier T1200. (iv) The gel results of carrier T1200 hybridized with 8, 16, 24, 32, 40 and 48 bricks.



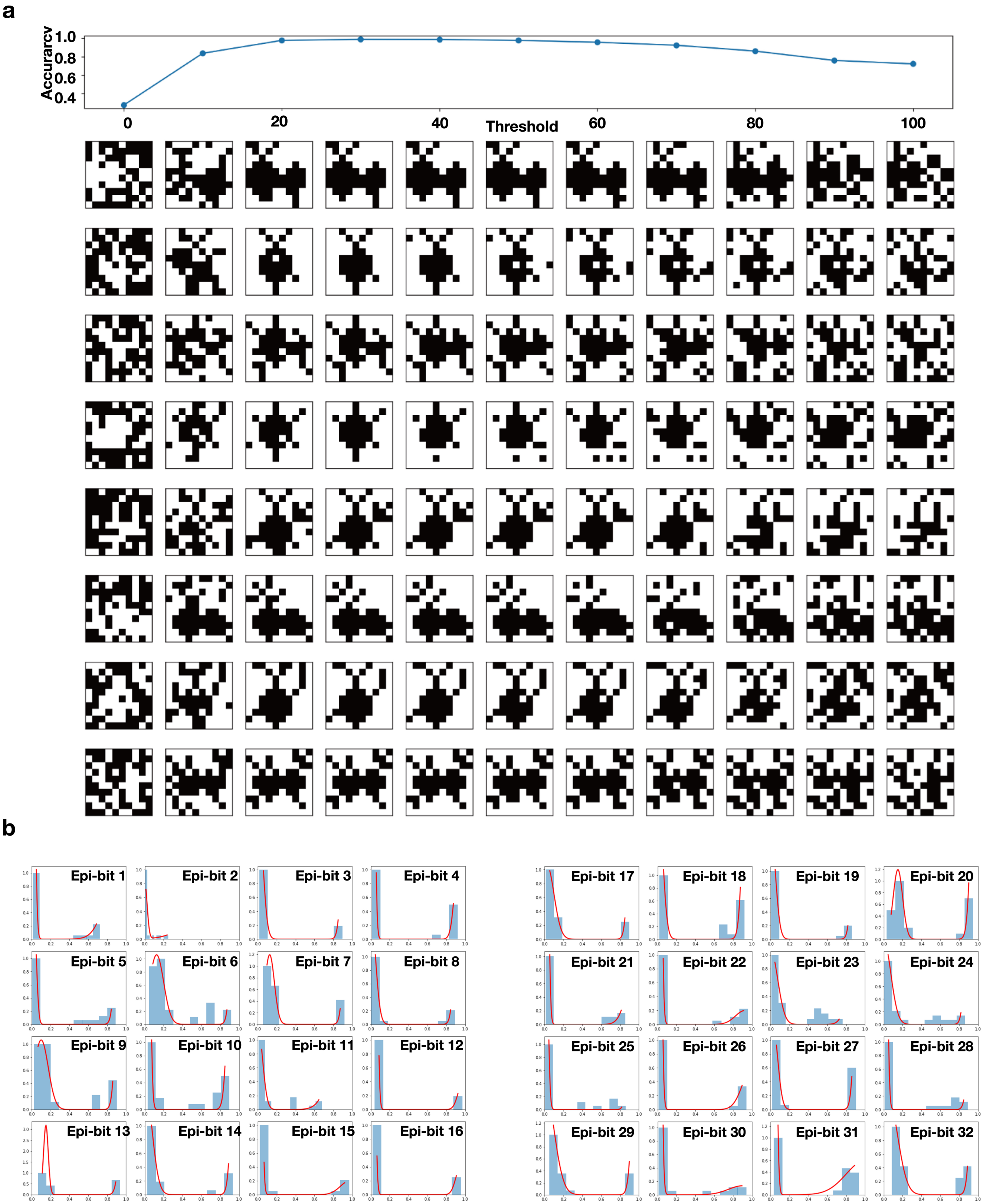
Extended Data Fig. 2 | The epi-bit writing tests. The epi-bit writing test on T1200 DNA carrier with 17 modification sites. The numbers of methylation sites were designed as one (117, 163, 209, 439, 485, 531), two (579–597, 603–621, 629–643, 826–830) and three (917-924-931) in the 24 nt bricks. **a**, The designed modification pattern. Solid and open ovals represent modified and non-modified CpG sites. **b**, The read coverages in nanopore sequencing along T1200. **c**, The methylation calling results from individual sequencing reads. **d**, The predicted methylation probabilities by megalodon. Red dots indicate the predesigned epi-bit 1 methylation sites. Box plots show the distribution of the data in bar plots (n = 187 for epi-bit 0, n = 17 for epi-bit 1). **e**, The durability of methylation storage under 25 °C and 95 °C for 1 h. **f**, Correlation between methylation calling results under the conditions of 25 °C and 95 °C.



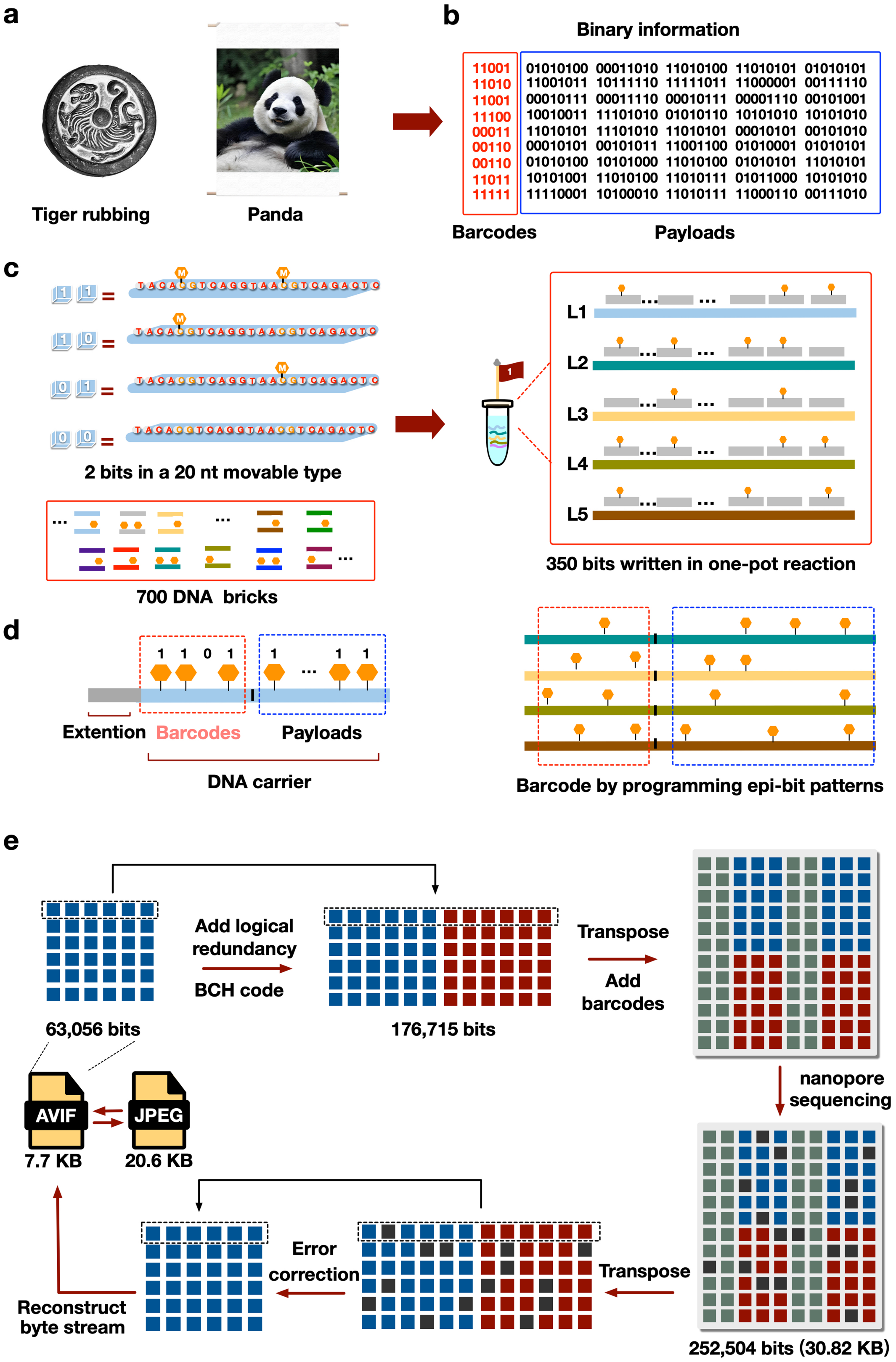
Extended Data Fig. 3 | The designs and experimental verification of the barcoded templates for parallel epi-bit storage. a, Ligation of barcodes and DNA templates to form barcoded DNA carriers. **b**, Electrophoretic gel verification of the barcoded templates and their modified forms. i, the 960 nt DNA carriers (from lanes 3 to 6) labeled with the specific barcodes by T4 ligation (Lane 1: 1,000 bp ladder). ii, purified 960 nt carriers without barcode (lane 4) and with barcode (lanes 5 and 6); iii, the gel results of DNMT1 transferred 960 nt carriers. Lane 1: barcoded 960 nt carriers; lane 2-3: 960 nt carrier hybridized with DNA bricks. **c**, Resolving the mixing ratios of three barcoded carriers by nanopore sequencing. Three differently barcoded 960 nt DNA carriers are mixed at varying concentrations for ratios of 1:1:1 (i), 1:1:48 (ii) and 10:100:1 (iii), respectively. Different colors of dots correspond to sequenced ratios at different sequencing coverages (200x; 500x; 1,000x; 2,500x and 5,000x).



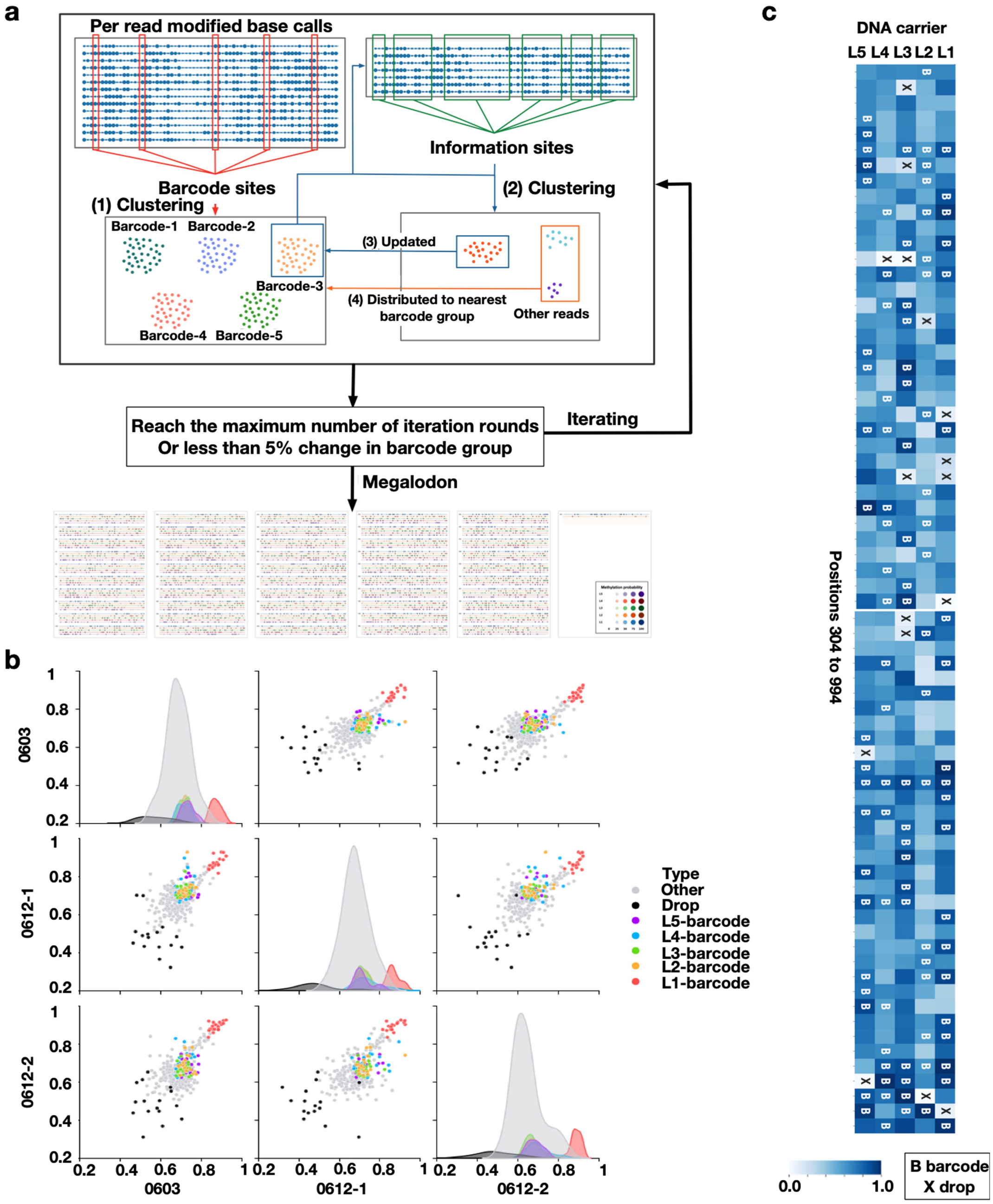
Extended Data Fig. 4 | The workflow and the cluster results of methylation information retrieval from one-pot nanopore sequencing. a, The methylation information retrieval workflow. We firstly used Megalodon to obtain the methylation information on individual sequencing reads (in the form of methylation probabilities). Next, we performed k-means clustering for all sequencing reads based on their called methylation patterns. After clustering, a barcode for each cluster were determined by majority voting of all reads within the cluster, and those with a nonconforming barcode were labeled incorrectly clustered reads. Next, only the correctly clustered reads were used to restore the methylation information for each cluster (solid grey arrows). The pink dashed line arrows show a pipeline in which methylation calling was performed without using the barcode information. This pipeline had a higher error rate and therefore was not adopted. Alternatively, reads could be grouped by their barcode sequences and then the methylation information could be called from reads within each group (solid red arrows). The original information was then retrieved based on the methylation information from the clusters. **b**, The clustering results of raw reads of 7 carriers (DNA carriers 1–6 and 25, left) and its density map (right). **c**, The numbers of clustered reads for each of the 7 carriers showing the consensus barcodes and the incorrectly clustered barcodes.



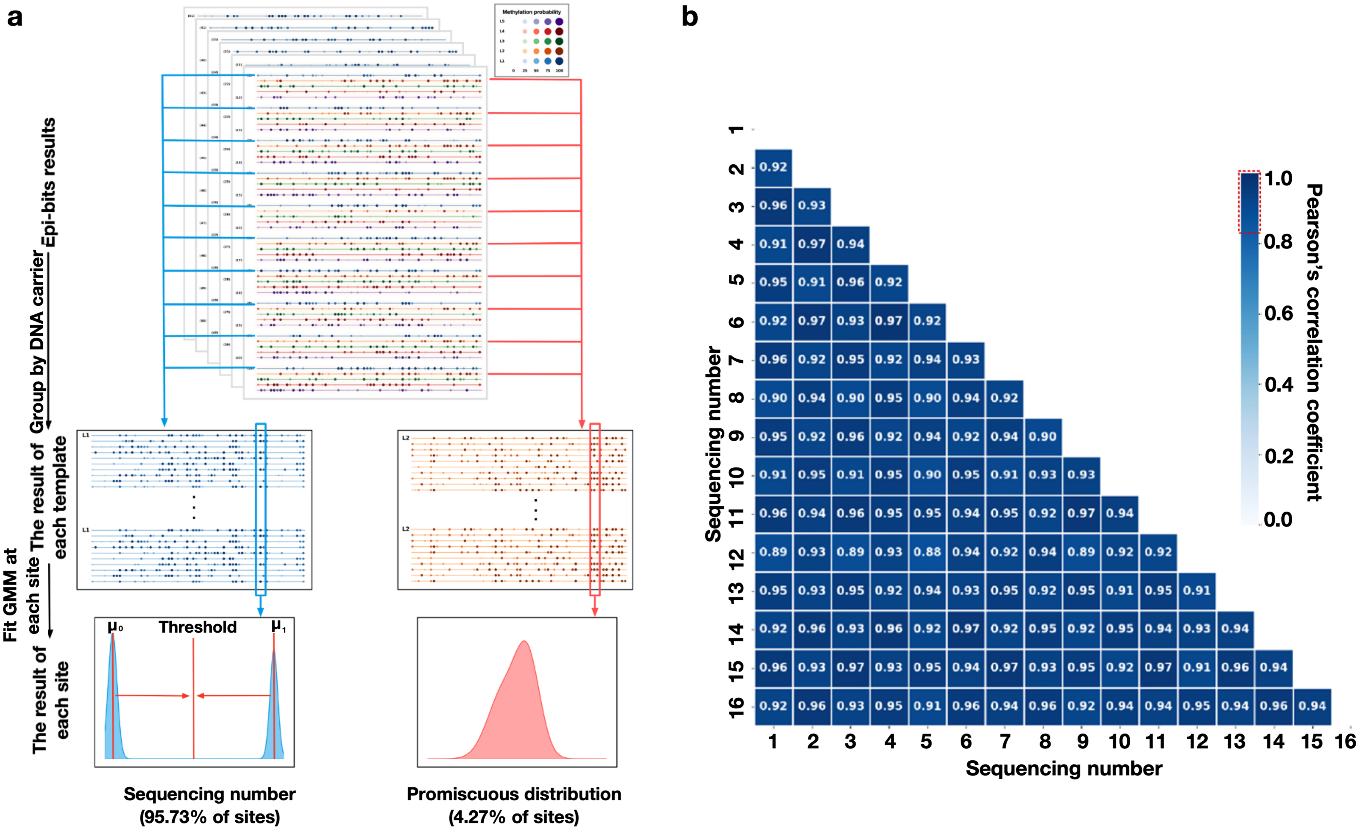
Extended Data Fig. 5 | Characterization of methylation information retrieval. a, Retrieved images under different thresholds of methylation calling. The recovery of 8 images (rows) were visualized under 11 different methylation calling thresholds from 0 to 100% (columns). The relationship between threshold and accuracy is shown in the top panel. **b**, The fitting results of 32 methylation sites on the DNA templates. For each methylation site, the results of methylation calling are collected (bar plots) and then used for fitting. The best fitting results (with minimal fitting loss) are illustrated as red lines. In each subplot, the x-axis shows the methylation probability, and the y-axis is the maximum-minimum normalized counts.



Extended Data Fig. 6 | The workflow of encoding, writing and decoding the epi-bit information. a, The original information. **b**, The original information was encoded to bit streams containing payloads and barcodes. **c**, Two epi-bit positions were designed in each 20 nt DNA brick as one movable type and 350 epi-bits were written on 5 different DNA carriers in one-pot reaction. **d**, Both barcodes and payloads were stored in DNA carriers as epi-bits. **e**, The workflow of encoding and decoding for the storage of the colored panda image.



Extended Data Fig. 7 | The workflow of dynamic clustering and the optimizations of epi-bit sites for large-scale epi-bit storage (the panda image). **a**, The dynamic clustering workflow to reduce the misclassified reads. First, the probability of methylation on the barcode sites was used as the basis for K-means clustering (Step 1). Second, reads in each barcode group were clustered based on the methylation probability of the information sites (Step 2). Third, adaptively selected the number of clusters depending on the silhouette score and then updated the barcode groups with the cluster closest to the current barcode in Euclidean space (Step 3). Meanwhile, other reads were distributed to the nearest barcode groups (Step 4). Finally, Megalodon were used to call the methylation probabilities from grouped reads. **b**, Per site accuracies between pairwise experiments for the optimizations of epi-bit sites. The accuracy of each epi-bit site in each sequencing experiment was calculated from the methylation calling results of single reads. Sites with accuracies less than 0.5 were dropped in the following experiments. The subplots on the diagonal are the kernel density estimates for each experiment. **c**, Heatmap of average accuracies at each site (columns) for each template L1-L5 (rows).



Extended Data Fig. 8 | The workflow of threshold selection and the correlations of the sequencing accuracies for large-scale epi-bit storage. **a**, The workflow of threshold selection. The thresholds of methylation calling on each epi-bit site were determined independently to avoid the DNMT1 context dependency. The methylation calling results were firstly grouped by DNA sequences of the carriers. Next, the methylation probabilities of each site on the carriers were fitted by a GMM (Gaussian mixture model). In the methylation results at each site, two peaks appeared on the fitted curve. In the fitted results, 95.73% sites followed clearly bipartite distributions, and the rest sites (4.27%) were indistinguishable for 0 s and 1 s. **b**, The correlations of the accuracies in 16 nanopore sequencing reactions in the experiments of the panda image. Here, the correlations are the Pearson's correlation coefficients for per-site accuracy in every pair of sequencing batches. Note that the single-read methylation probabilities were used for the site-specific threshold determination based on GMM.