nature communications

Article

https://doi.org/10.1038/s41467-025-61264-5

Random access and semantic search in DNA data storage enabled by Cas9 and machine-guided design

Received: 24 January 2025

Accepted: 18 June 2025

Published online: 10 July 2025

Check for updates

Carina Imburgia[1,4], Lee Organick[1,4], Karen Zhang[1,4], Nicolas Cardozo[1,4], Jeff McBride[1], Callista Bee[1], Delaney Wilde[1], Gwendolin Roote[1], Sophia Jorgensen[1], David Ward[1], Charlie Anderson[1], Karin Strauss[2], Luis Ceze[1] & Jeff Nivala[1,3]

DNA is a promising medium for digital data storage due to its exceptional data density and longevity. Practical DNA-based storage systems require selective data retrieval to minimize decoding time and costs. In this work, we introduce CRISPR-Cas9 as a user-friendly tool for multiplexed, low-latency molecular data extraction. We first present a one-pot, multiplexed random access method in which specific data files are selectively cleaved using a CRISPR-Cas9 addressing system and then sequenced via nanopore technology. This approach was validated on a pool of 1.6 million DNA sequences, comprising 25 unique data files. We then developed a molecular similarity-search approach combining machine learning with Cas9-based retrieval. Using a deep neural network, we mapped a database of 1.74 million images into a reduced-dimensional embedding, encoding each embedding as a Cas9 target sequence. These target sequences act as molecular addresses, capturing clusters of semantically related images. By leveraging Cas9's off-target cleavage activity, query sequences cleave both exact and closely related targets, enabling high-fidelity retrieval of molecular addresses corresponding to in silico image clusters similar to the query. These approaches move towards addressing key challenges in molecular data retrieval by offering simplified, rapid isothermal protocols and new DNA data access capabilities.

DNA is a naturally robust information storage molecule that features ultrahigh information density (17 exabytes/gram) and long-term chemical stability, capable of retaining its integrity for thousands to millions of years[1–6]. These features have caused DNA to be widely explored as a storage medium for digital data[3,7–14]. As DNA synthesis and sequencing costs continue to decrease, DNA will likely only become an increasingly attractive storage medium[15–17]. Furthermore, recent work has explored scaling up DNA data storage[1,11,12,14,18] and developing end-to-end workflows that encompass encoding, synthesis, retrieval, sequencing, and decoding of digital data stored in DNA[7,13,19,20]. In early DNA data storage architectures, targeting specific files required sequencing the entire DNA library[3,8,10]. Recently, efforts to reduce costs and increase resource efficiency have motivated emphasis on developing increasingly efficient retrieval methods in DNA data storage systems, which read specific files or subsets of data using techniques such as random access or similarity search.

One of the most common random access strategies is to use conventional PCR-based enrichment[7,9,11,13,18]. Although performing the

[1]University of Washington, Paul G. Allen School of Computer Science and Engineering, Seattle, USA. [2]Microsoft Research, Redmond, USA. [3]University of Washington, Molecular Engineering and Sciences Institute, Seattle, USA. [4]These authors contributed equally: Carina Imburgia, Lee Organick, Karen Zhang, Nicolas Cardozo. e-mail: jmdn@cs.washington.edu

PCR retrieval step is relatively straightforward, there exist prohibitive limitations. The end-to-end protocol is time-consuming and requires careful design and validation of primers to avoid molecular crosstalk with data payloads. Furthermore, multiplex retrieval (i.e., targeting multiple files in a single reaction) is challenging in large pools[11,13,18,21]. These issues worsen as databases are scaled up, restricting the upper limit of database size. To circumvent these challenges, some designs use physical separation of unique files. The DENSE (DNA Enrichment and Nested SEparation) storage system combines physical separation with nested addressing[14]. However, the magnetic bead extraction central to this method makes it complex and multi-step and does not demonstrate multiplexed access. Other retrieval techniques physically encapsulate files to achieve separation[22,23]. Banal et al. demonstrated fluorescent-assisted sorting (FAS) and, through tagging capsules with addresses or metadata keys, enabling operations such as boolean search and multiplexed access. Bogels et al. expanded PCR-based access with the introduction of controlled encapsulation. These methods achieve multiplexed and repeated random accesses to DNA databases by mitigating molecular crosstalk, but are challenging to implement and present new challenges to scaling systems. For example, FAS is limited by the number of orthogonal fluorophores available per sorting step. Scaling this method would require multiple rounds of sorting, impacting efficiency. Furthermore, the encapsulation method currently takes days[22] and significantly reduces data density and necessitates more complex library preparation steps, impacting overall scalability. Given these existing challenges, a random access technique that provides a simple protocol with multiplex access capabilities is highly desirable.

Similarly, there is a need to expand database retrieval beyond distinct file lookup. A well-curated DNA database might assign metadata or a unique key for precise lookup for each item. This requires the user know the exact metadata tags or keys of items they wish to target. Modern search platforms do not assume the user knows these identifiers or the content of the data stored. Content-based similarity search integrates data retrieval with compute abilities to target data interrelated to user-specified queries, e.g., reverse image search. Recently, similarity search was successfully performed in DNA by leveraging DNA hybridization[24]. One advantage of using DNA rather than traditional silicon-based computers is that the query molecules are able to diffuse in solution and interact with all items in the database in parallel, which could significantly improve search times at large scale. However, the previous hybridization-based similarity search work requires a significant and lengthy temperature gradient to perform search and a protocol that runs over 24 h.

The CRISPR-Cas9 system, known for its sequence-specific, programmable targeting, presents interesting potential for DNA data storage and retrieval. Prior work demonstrates its use as a write mechanism in cellular DNA[25,26] and content-based search capabilities through activation of a fluorescent readout when a targeted key is bound[27]. Here, we present CRISPR-Cas9-based random access (C9RA) and similarity search (C9SS) approaches for DNA data storage and pair each with nanopore sequencing. The Cas9 system's precision, derived from its adaptive immunity origins in bacteria, underpins our C9RA approach. This specificity facilitates the specific addressing of DNA data files, allowing for the selection of distinct files from a large DNA pool for nanopore sequencing-based readout. Conversely, the system's off-target cleavage activity, often viewed as a limitation, is leveraged in our C9SS approach. There are numerous in vitro and in vivo high-throughput studies that estimate wtCas9 off-target activity[28–31]. To harness this characteristic for similarity searches, we integrated a Cas9 cleavage prediction model[32] from one of these studies into a deep learning-based encoder. This integration allowed us to encode a database of 1.74 million images with semantic addresses, along with gRNA query sequences representing the visual attributes of five images. The retrieved addresses correspond to clusters containing the images most similar to the query, enabling efficient and accurate similarity-based lookups. Our resulting method for similarity search can be applied to any dataset capable of reduction to distinguishable semantic embeddings. While it trades some retrieval precision compared to its hybridization-based predecessor, Cas9SS offers a simple, rapid protocol for content-based retrieval. In essence, the dual capabilities of Cas9 offer potential avenues to address distinct data retrieval challenges in DNA data storage.

Results

Single and multiplex random access

Cas9-based targeting systems have been previously employed in genomics applications to selectively enrich specific chromosomal DNA regions for nanopore sequencing[33]. This involves a Cas9/gRNA ribonucleoprotein complex (RNP) that cleaves genomic DNA at regions of interest, exposing 5′ phosphates. These exposed ends then allow for selective ligation of nanopore sequencing adapters, preparing only the selected regions for sequencing. Adapting this for DNA data storage, we utilized a DNA pool encoding 25 batches of digital image files from the Memories in DNA project[34]. The payload strands were designed with file-specific Cas9 target sites (Fig. 1A), and the overall design was tailored to integrate with the Rolling Circle Amplification to Concatemeric Consensus (R2C2) method[35]. This method generates long concatemeric repeats of the DNA, which are nanopore sequenced and algorithmically aligned to produce a high-accuracy single-molecule consensus sequence.

For our C9RA approach (Fig. 1), the payload strands comprised three components: 1) A 125nt data payload region; 2) A 20nt Cas9 file address overlapping with universal forward primer sites, with the last 7nt being file-specific; and 3) A 20nt universal reverse primer site. These primers enabled PCR amplification of the entire ssDNA pool after array-based synthesis of our 1.6 million strand DNA library, which was divided across 25 files (approximately 60K strands per file).

After synthesis of the ssDNA pool, we converted it to dsDNA using PCR with the universal primer sequences. These PCR products were subsequently circularized through isothermal assembly[36] into a 430 bp splint vector. Rolling Circle Amplification (RCA) amplification was used to improve nanopore sequencing quality and downstream alignment. Initially, long concatemer strands (> 10kb, Supplementary Fig. 1) were produced and then dephosphorylated in line with the R2C2 protocol. While this method introduces more preparation steps compared to PCR-based random access techniques, it's crucial to highlight that these steps are "offline". This means they can be executed post-synthesis but before the DNA library is stored.

Following offline preparation, we initiated a preliminary assessment of our C9RA approach. We first chose File ID 10 to test the targeting of a single file (see "Methods"). During this test, a portion of the DNA pool underwent a 1-min incubation with the File 10-specific-RNP complex at 37 °C. Sequencing adapters were ligated to cut strands and read with a MinION nanopore sequencer. From the obtained data, we generated single-molecule payload consensus sequences for each read. To evaluate C9RA enrichment, we calculated an enrichment score (ES) for each file. The ES measures the ratio of reads for each of the 25 files after random access compared to their distribution in the original DNA pool (found in Fig. 2A). Detailed ES methodology is provided in Supplementary Section 2.

File 10 exhibited the highest enrichment. Sequencing data from this experiment showed a two orders of magnitude enrichment for File 10, with an decreased average concatemer length compared to the other files, suggesting Cas9 cutting (Supplementary Fig. 2i, ii). Summary plots further highlighted these findings (Fig. 2B, C).

We then evaluated the C9RA system's multiplex capability by simultaneously targeting Files 2, 13, and 24 (Fig. 3). Multiple-file targeting allowed for an incubation time of 15 min without the consequence of off-target cutting. After subsequent sequencing, only

https://doi.org/10.1038/s41467-025-61264-5

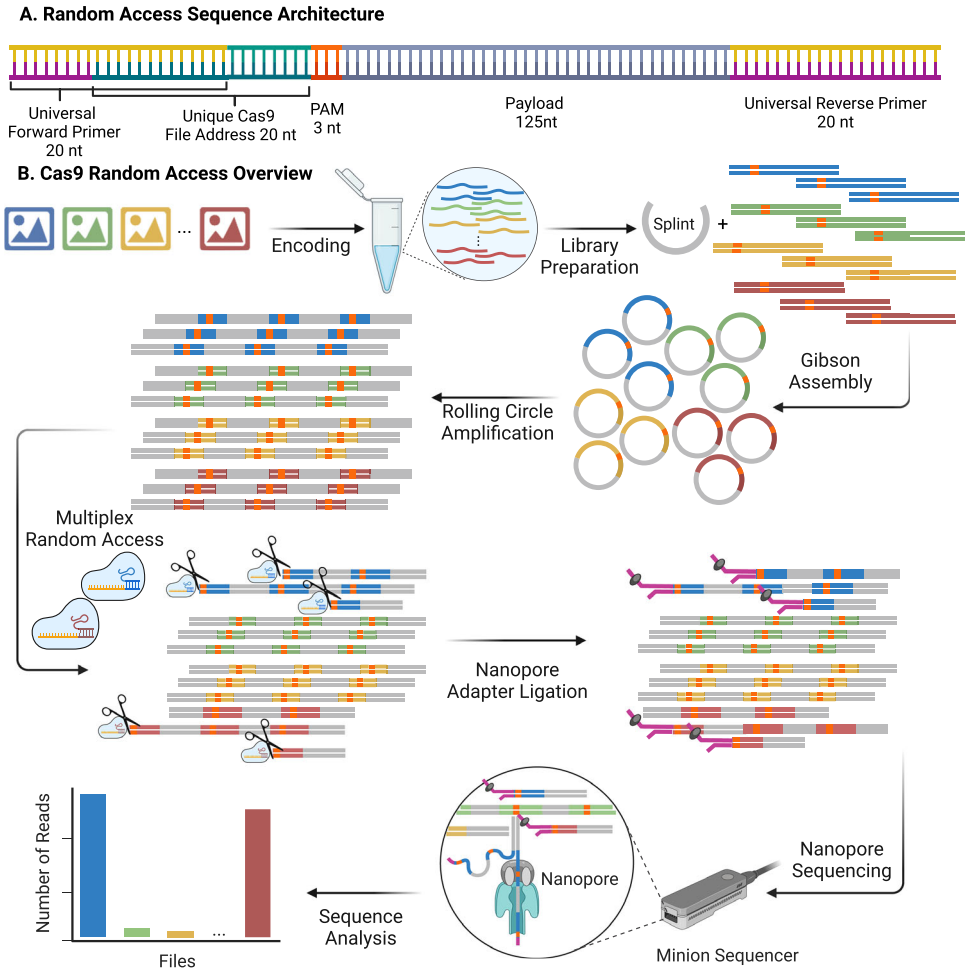


Fig. 1 | Schematic of Cas9 random access workflow. A Each 165 nt file has a payload and address that contains a unique Cas9 target site including a PAM. On either end of the payload are universal primers which hybridize to the ends of a universal splint during Gibson assembly. The Cas9 target site and universal forward primer share 13 nt. **B** Digital files are encoded into DNA sequences with the architecture shown in **A**. The sequences are circularized and amplified via rolling circle amplification (RCA) resulting in linear, high molecular weight strands. The linear DNA can now be stored for future access. At retrieval, specific file(s) from a set of 25 are targeted with their complementary Cas9-RNPs, which cut the long DNA strands at file address sites. Adapters for nanopore sequencing are ligated only at the cut sites where there is an exposed phosphate group, thus only these files with the ligated adapter are sequenced and retrieved. Created with BioRender.com.

these files showed substantial enrichment, with scores at least an order of magnitude higher than other files (Fig. 3A). Expanding our test, we targeted all even-numbered file IDs (12 in total) and then 20 distinct files. On average, targeted files exhibited 100-fold higher enrichment scores (Fig. 3B, C). Across all multiplex tests, most targeted files outperform untargeted ones; however, as the number of targeted files increases, we begin to see an overlap between the enrichment scores of the two groups (Fig. 3D).

Similarity search encoding and model performance

Having established the efficacy and multiplexing capability of our C9RA approach, we sought to further demonstrate the potential of Cas9 in a different data retrieval context: semantic search. While the random access approach focuses on precise targeting of a known key, similarity search aims to retrieve data based on likeness or resemblance. Central to the development of the Cas9-based similarity search (C9SS) approach was to encode a large image database into DNA sequences (architecture in Fig. 4A) where images with similar features are more likely to have Cas9 guide/target binding. This ensures that when an image is encoded as a query (represented by the gRNA strand) and matches a similar image in the database (the semantically-addressed target DNA sequence), the query's Cas9-gRNA complex cleaves the target DNA sequence for subsequent readout via selective nanopore sequencing (Fig. 4B). The encoding process begins by leveraging the VGG-16 image classification model's FC2 layer to transform each image into a 4096-feature vector[37]. Following this transformation, a neural network-referred to as the sequence encoder or encoder model-is trained to map these image feature vectors onto DNA sequences. The predictability of interactions between the query and target, specifically the likelihood or rate at which the gRNA will cleave a target sequence, is pivotal for the training and evaluation of the sequence encoder. Our encoder integrates a model that predicts Cas9 cleavage rates between sequence pairs[32]. The Euclidean Distance between the feature vectors of two images allows image pairs to be categorized as similar or dissimilar to each other. We selected a distance threshold of 75, where ≤ 75 is considered similar and > 75 dissimilar. A detailed outline of the encoding procedure is available in the Methods section, with a visual summary presented in Figure 4C.

Determining the optimal cleavage rate at which a target is considered retrieved (referred to as a yield threshold) is a non-trivial task. An ideal encoder model with an ideal threshold would retrieve all similar and no dissimilar images. However, constraints on encoding (discussed in depth in Supplementary Section 7 and 8) and the molecular mechanisms of the retrieval reaction make this unfeasible. The

https://doi.org/10.1038/s41467-025-61264-5

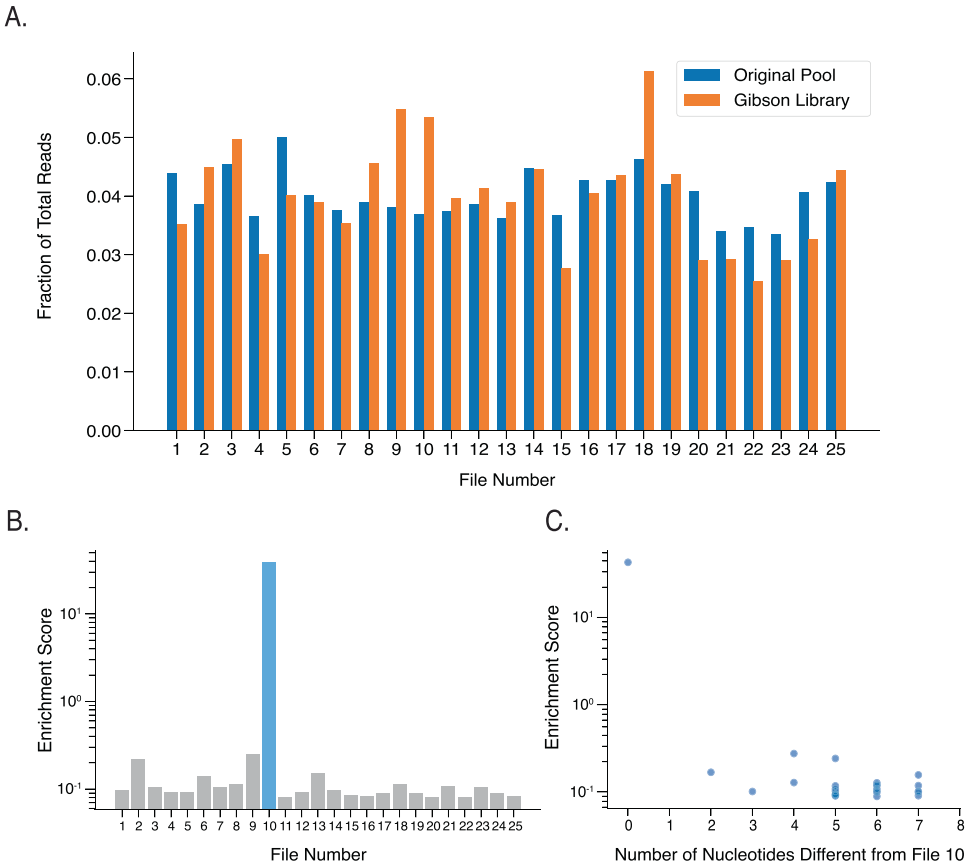


Fig. 2 | Single-file enrichment experiment. A Distribution of files in the baseline pool before and after Gibson Assembly. Sequencing results show each file is relatively represented within the unenriched pool. **B** Summary plot showing the distribution of enrichment scores for targeted File 10 vs the untargeted files in the same experiment. A 1 µM concentration of gRNA was used. After a one-minute incubation with Cas9-RNPs, File 10 was enriched two orders of magnitude over the other files, and no untargeted files were enriched. **C** Files with Cas9 gRNA target sequences more similar (same nucleotides in same positions) to that of File 10 are more likely to be co-enriched when File 10 is targeted.

encoder model's performance changes depending on the threshold used. To examine the accuracy of our encoder model as a function of yield threshold, we simulated retrieval on 50,000 query-target pairs, encoded from random images in the OpenImages database. Half were pairs of two similar images, and the other half were two dissimilar images. Figure 5A shows the proportion of targets, similar (y-axis) and dissimilar (x-axis) to their paired query, that are predicted as retrieved as the yield threshold decreases. Random performance is represented by the grey dashed line. To compare performance between encoder models, the area under the receiver operator characteristic curve (*AUC*) was calculated. We trained three encoder models and used the model with the greatest *AUC* to encode our full data set. A threshold of 0.1 was used for the remainder of the analyses to maximize retrieval of similar images and minimize retrieval of dissimilar images.

The best-performing encoder model was used to encode a database of 1.74 million images unseen during training. Encoding resulted in 457 unique sequences (a complete list can be found in the GitHub repository listed in the Data Availability Statement). The number of images that map to one of the 457 unique sequences ranges from one to one hundred fifty thousand and is detailed in Supplementary Section 5. All 457 sequences were synthesized as DNA oligomers according to the architecture shown in Figure 4A and used as a database for laboratory experiments (see Methods for implementation details).

Similarity search implementation

The C9SS method was implemented in the wet lab with five distinct image queries. Each query was encoded as a gRNA sequence using the same model that encoded the database. Cas9-RNPs with encoded gRNA sequences were then prepared to perform retrieval within our DNA-encoded database. In parallel, the entire database was amplified and a portion was sequenced for a comparative baseline. For each retrieval experiment, a small portion of the amplified baseline pool was dephosphorylated and then combined with the Cas9-RNPs of one of the image queries. Consistent with our C9RA approach, Cas9 cleavage of the data-encoded DNA strands exposes a 5′ phosphate group, which ligate to nanopore adapters for selective readout during sequencing. Sequencing data was used in combination with experimental simulations to evaluate and understand the performance and mechanism of C9SS. We highlight two separate query images (a picture of a cat, and the iconic picture of Bigfoot from frame 352 of the 1960s-era Patterson-Gimlin film) to discuss our results (Fig. 5B). The complete results can be found in Supplementary Fig. 6.

The extent that simulated results are predictive of the in vitro enrichment of target strands informs the accuracy of our retrieval architecture at the DNA sequence level. We evaluated the cleavage predictor relationship to observed enrichment. Enrichment scores for each of the 457 targets were calculated by comparing read count ratios of retrieved sequences to the baseline (details in Supplementary Section 2). The linear and monotonic correlation between the predicted score and target enrichment from each query is observed in Fig. 5C, D. Furthermore, enrichment scores were calculated from a baseline sequencing run unused in other analyses and our comparative baseline to represent an unqueried or randomly enriched database. Importantly, the predicted cleavage activity for both the cat and bigfoot

https://doi.org/10.1038/s41467-025-61264-5

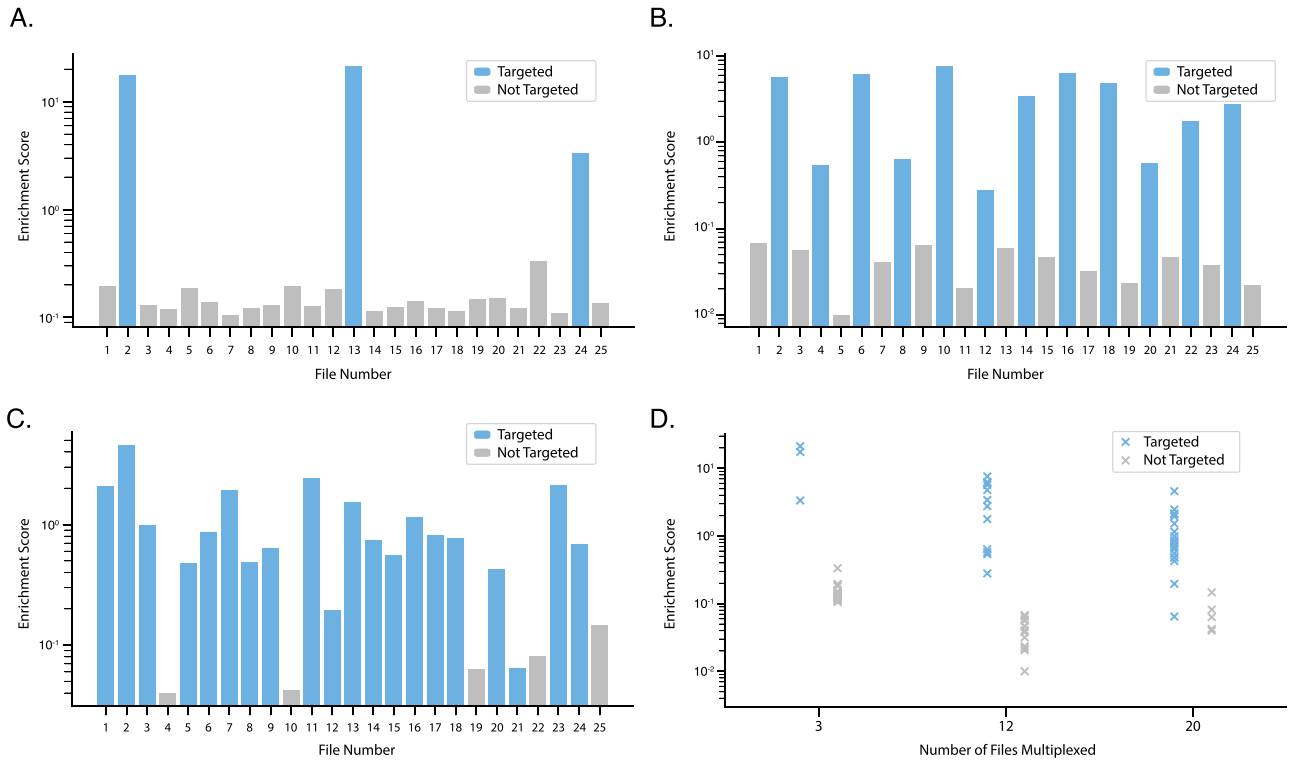


Fig. 3 | Multiplexed enrichment experiments. A Enrichment score of each file after targeting 3 files in a 25-file pool. 3.33 µM of each targeted file's gRNA was incubated in a Cas9-cleavage reaction for 15 min prior to sequencing. The targeted files were enriched to at least one order of magnitude greater than the untargeted files. **B** Enrichment score of each file after targeting 12 files in a 25-file pool. 0.83 µM of each targeted file's gRNA was incubated in a Cas9-cleavage reaction for 1 min prior to sequencing. All targeted files had a higher enrichment score than any untargeted files. **C** Enrichment score of each file after targeting 20 files in a 25-file pool. 0.5 µM of each targeted file's gRNA was incubated in a Cas9-cleavage reaction for 15 min prior to sequencing. All but one targeted file (File 21) had a higher enrichment score than any untargeted files, resulting in a single false negative. Generally, the difference in enrichment between targeted and untargeted files decreases for larger multiplexed sets. **D** Summary plot showing distribution of enrichment scores for targeted vs untargeted files from panels **A**–**C**.

image queries correlated with target enrichment scores ($r = 0.65$, $\rho = 0.68$ and $r = 0.60$, $\rho = 0.62$, respectively).

We next evaluated how effectively our model's encoding of the database facilitates the retrieval of information that is semantically similar to the query images. To do this, we quantified the enrichment of images similar to the query and depletion of dissimilar images for each of the two image search experiments and also compare it to simulated results. The heat maps in figure 5E report simulated and actual retrieval results for all 1.74 million images, ordered by their Euclidean distance to a query and grouped by ranges of Euclidean distance and sequencing yield. Each cell is colored by the proportion of images within a row that fall into its yield range. Targets are classified into four categories: similar (≤ 75) or dissimilar (> 75) and retrieved (> 0.1) or not retrieved (≤ 0.1) as distinguished by the dotted lines. Simulated results suggested that both queries will retrieve the most similar images in the database (as indicated by the high heat density in the upper right quadrant), but also a proportion of dissimilar. The experimental results for similar images align well with the simulated results. The cat and Bigfoot query retrieved 98% and 62% of similar images, respectively. A subset of the two most similar and two least similar images in each category are shown in figure 5F. Observation of these samples provides us with some insight into performance variation by query. The cat query, while represented in a large set of cat images in the database, has highly distinguishable features that increase its distance from other images. Consequently, we see high performance in the retrieval of similar images. While it still enriched the majority of similar images, the search for Bigfoot remained more elusive as its image is more broadly generalized to many items in the database.

Discussion

In summary, we introduced CRISPR-Cas9 as a versatile tool for both random access and similarity search in DNA data storage. Our C9RA architecture allows for multiplexed file targeting using the CRISPR-Cas9 system and nanopore sequencing. This method offers several advantages, including isothermal enrichment, expanded multiplexability, faster reaction times, and reduced time-to-retrieval (from hours to seconds). We have successfully multiplexed random access to 10 files and showed 20-file multiplexed targeting enriches almost all targeted files. This is on a similar order of magnitude to Bogels et al., who demonstrated multiplexed amplification of 25 files[23]. Further exploration into the limits of multiplexing C9RA is left for future work. Additionally, we aim to achieve more densely-packed addressing of file sequences and anticipate further simplifying the automation of DNA data storage pipelines. However, a potential limitation of this method is the destruction of targeted strands after readout. This issue could be mitigated by using catalytically inactive Cas9 (dCas9) for pull-down-based enrichment during sequencing.

Our C9SS similarity search architecture enables the retrieval of file addresses that point to content stored in silico that is similar to a query, without the need for preexisting exact knowledge of a database's contents, aside from access to the encoder. This system also features an isothermal search protocol, a rapid reaction time of 30 s, and a simpler implementation than previous methods. It also proves more energy-efficient and faster compared to hybridization-based approaches. A fast prep and reaction rate reduce thermocycler time from about 24 h to tens of seconds, thereby consuming significantly less energy. For example, assuming a thermocycler operating at 100 W, the energy consumption drops from approximately 2.4 kWh (for 24 h)

https://doi.org/10.1038/s41467-025-61264-5

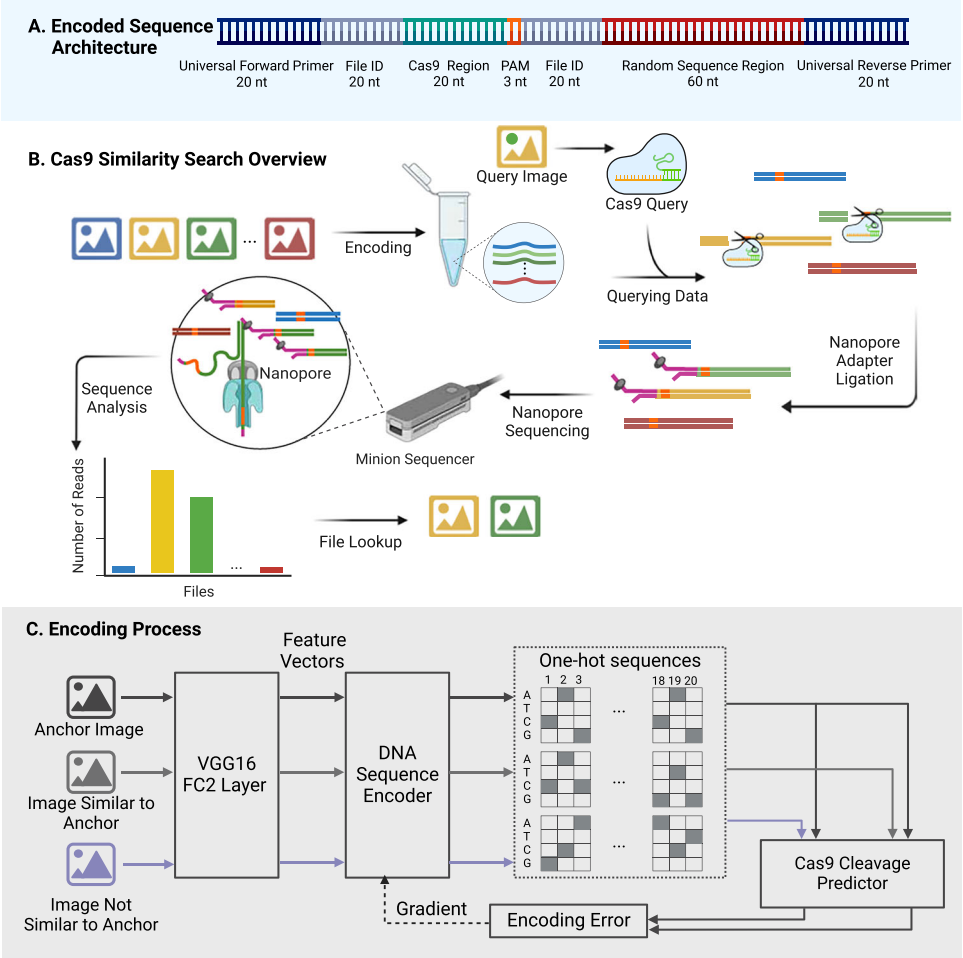


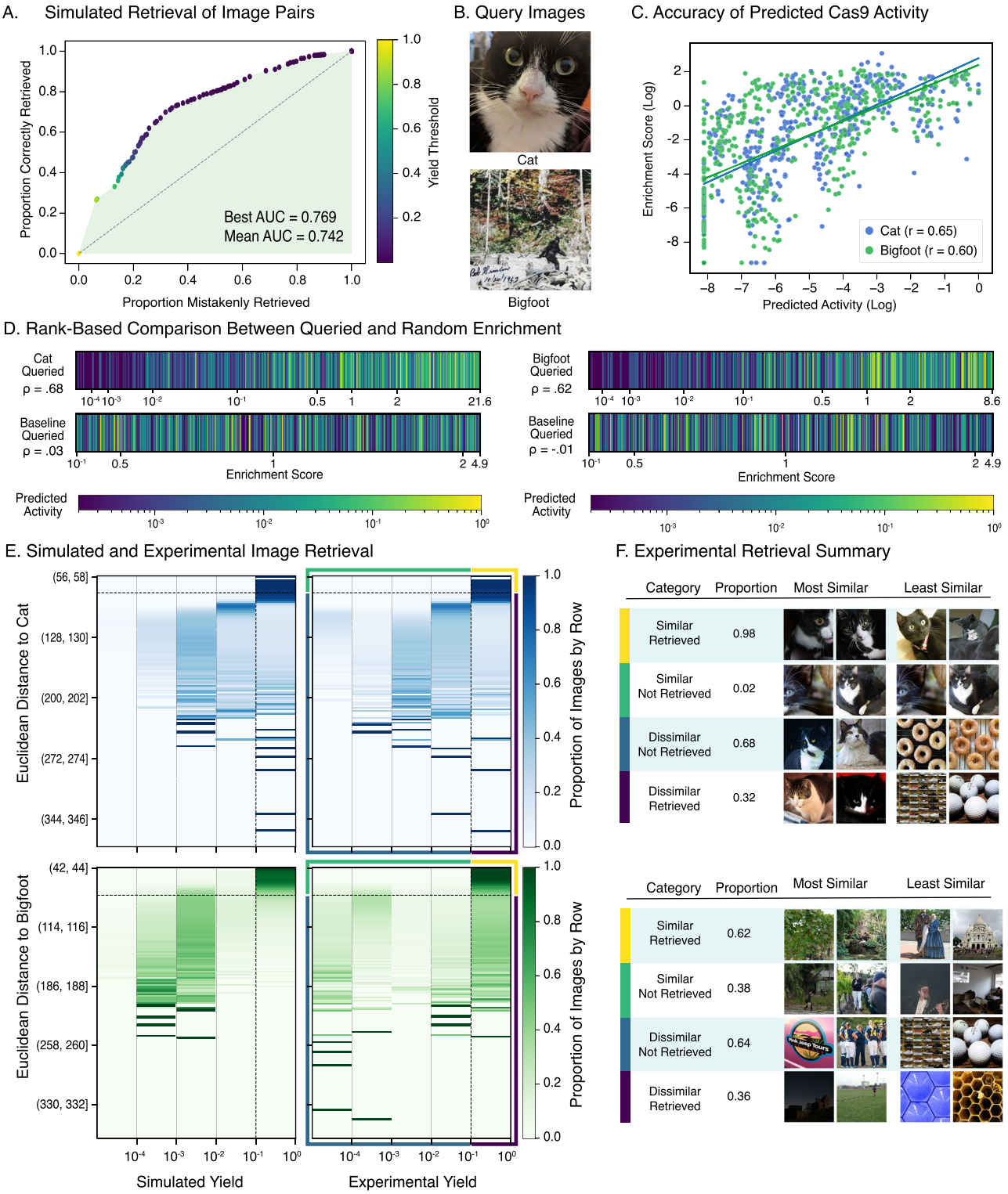
Fig. 4 | Schematic of Cas9 similarity search workflow. A The sequence design after encoding. Each File ID corresponds to one item from the database, but all items in the database have the same forward and reverse primer sequences (see Supplementary Section 9). Each sequence has a Cas9 region, 20 nt long, followed by a 3 nt PAM sequence (TGG). A 60 nt random sequence region is included to improve alignment. **B** Images are represented by a DNA sequence assigned during encoding. A query image is also encoded to generate the complementary Cas9-RNPs that cut the similar sequences in the database. Adapters for nanopore sequencing are ligated only at the cut sites where there is an exposed phosphate group, thus only these sequences with the ligated adapter are sequenced and retrieved. The retrieved File IDs can then be used to look up the associated images. **C** Batches of image triplets from the training data set are used to train the encoder model. Each triplet consists of an 'anchor' image, one image similar to, and one dissimilar to the anchor. Images are transformed to feature vectors using the FC2 layer of the VGG16 image classification model. The feature vector acts as a summary of the image. The shorter the Euclidean distance between a pair of feature vectors, the more similar the corresponding images are. All three feature vectors from a triplet are passed through two fully connected layers (the 'Encoder'), resulting in one-hot vectors for each image, which represent an encoded DNA sequence. A Cas9 cleavage model then operates on the sequences to predict the Cas9 activation rates between the anchor (query) and the similar and dissimilar images (targets). The activation rates are used to compute a loss, which is back-propagated to train the encoder to maximize the activation on similar pairs while minimizing it on dissimilar pairs. Created with BioRender.com.

to just a few watt-hours for a reaction that takes tens of seconds. C9SS performance hinges on the predictive accuracy of the cleavage rate simulator and the encoder's performance. Compared to the previous hybridization-based similarity search work, our Cas9-based results show a higher retrieval rate of dissimilar images, which we hypothesize to be, in part, related to Cas9's high specificity for sequence recognition (detailed in Supplementary Sections 7 and 8). Cas9 cleavage rates decrease rapidly from sequence changes between the query and target, particularly in the PAM proximal region[32], and often by as much as a factor of 10^{-3} after just 2 mutations (Supplementary Fig. 7). Consequently, the model's mutable sequence space is reduced, impacting attempts to represent more complex relationships across images during encoding. For example, if A is similar to B and C, but B and C are not similar, the encoder has limited space to make sequence changes between B and C while also preserving the similarity both have to A.

However, it is possible that wtCas9 has more encoding space than initially predicted. The Cas9 cleavage predictor is a biophysical model, developed using datasets that limited query-target mismatches to two and measured activity rates. This constraint likely accounts in part for the discrepancy between our predicted and experimental results, as our datasets include query-target pairs with more mismatches. Although we observed a moderate correlation between simulated and experimental results in Fig. 5c, d, the simulator often overestimated the reduction in Cas9 activity due to mismatches.

Although not explored here, several concrete paths to improving the C9SS system exist. For example, in order to remain consistent with our previously developed similarity search architecture[24], we chose to use the feature vectors extracted from the VGG-16 model (see our Data Availability section). However, it is possible that we could see performance improvements during the image classification step and encoder training with the use of more powerful models, which offer accuracy gains using fewer parameters[38,39]. Another strategy to improve C9SS performance could involve using a Cas variant with increased off-

https://doi.org/10.1038/s41467-025-61264-5



target activity for more flexible similarity searches, though such variants are not yet commercially available to our knowledge. This approach might, however, compromise the precision of the C9RA system. A combined approach of C9SS and C9RA in a hierarchical retrieval framework could also be beneficial, utilizing natural Cas9's specificity for accurate random access and less stringent Cas variants for less exact semantic searches. Both methods significantly enhance energy efficiency and retrieval speed over existing molecular methods, and provide a promising direction for future developments in molecular information processing.

Methods

Random access DNA library synthesis and amplification

The DNA pool containing the 25 files was synthesized by Twist Bioscience into 1.6 million DNA strands. The 430 bp splint sequence was obtained via PCR from pCDB180 (https://www.addgene.org/80677/). The DNA pool (payload) and the splint were amplified with 2x KAPA Master Mix with 10 µM reverse and forward primers (95 °C for 3 min; then 11 (payload) and 15 (splint) cycles of: 98 °C for 20 s, 62 °C for 15 s, and 72 °C for 30 s; followed by final 72 °C for 30 s). DNA was purified using the QIAGEN QIAquick PCR Purification Kit protocol.

Fig. 5 | Evaluating simulated and experimental similarity search. A 50,000 balanced image pairs are generated, i.e., half the pairs are of two similar images (Euclidean distance ≤ 75), and half are of two dissimilar (Euclidean distance > 75). Each data point represents the proportion of the 50,000 image pairs that was correctly or mistakenly retrieved using a given yield threshold, as denoted by the color bar. Yield threshold corresponds to the Cas9 cleavage rate that must occur on the target to be classified as a recall. The dashed line represents random recall. The area under the curve (AUC) is shaded green and is a quantitative metric for model performance. 1.0 is the maximum possible AUC value. **B** The two images used for retrieval experiments. **C** Enrichment score as a function of predicted activity. Each point represents a sequence in the database and a log-scaled relationship is plotted for both experiments. **D** Enrichment scores were calculated from sequencing results from either our queried database or a baseline unused in previous analyses (representing an unqueried database) to evaluate the predictive accuracy of our simulator. All target sequences are ordered by their enrichment score (x-axis) and colored by predicted cleavage from the query. **E** Heat maps of simulated and experimental retrieval results from the Cat (blue) and Bigfoot (green) queries. Images are binned by Euclidean distance and a normalized simulated yield or enrichment score. Color indicates the proportion of total images in a row that fall into each yield bin. Images are partitioned into four categories (similar or dissimilar and retrieved or not retrieved), distinguished by the dotted lines. **F** Summary of the proportion of similar and dissimilar images and the two most and least similar images in each category.

Random access library preparation: assembly and amplification of circularized DNA

Gibson assembly and Rolling Circle Amplification (RCA) of the splint and payload were prepared as described in figure 1. In the Gibson Assembly, 200 ng of both splint and payload were combined with 2x NEBuilder HiFi DNA Assembly Master Mix (NEB) and water, and incubated for 60 min for 55 °C. The Gibson product was digested with 1 µL each of 1:10 Exonuclease III, Lambda Exonuclease, and Exonuclease I, all from NEB. The circularized DNA was extracted using AMPure XP Beads at a ratio of 1.6 beads:1 sample and eluted in 25 µL water.

In the Rolling Circle Amplification, 10 µL aliquots of circularized DNA were amplified in 50 µL reactions with 5 µL of 10x Phi29 buffer (NEB), 2.5 µL of 2.5 mM dNTPs, 2.5 µL of 100 µM random hexamers (Thermo), 1 µL of Phi29 polymerase (NEB), and the volume was adjusted with water. Reactions were incubated overnight at 30 °C. RCA product was extracted using AMPure XP Beads at a ratio of 0.5 beads:1 sample. DNA was debranched and eluted by adding 10 µL NEB buffer 2, 2.5 µL T7 Endonuclease and 90 µL of water to the beads which were then incubated on a thermal shaker at 37 °C for 1 hr. The supernatant from the beads was collected on magnets, and the DNA in it was extracted again using 0.5 AMPure VP beads:1 sample and eluted in 15 µL water.

Random access Cas9 enrichment

Cas9 enrichment was accomplished as per Oxford Nanopore Technology (ONT)'s protocol for Cas-mediated PCR-free enrichment. Cas9 ribonucleoprotein complexes (RNPs) were prepared by combining 3 µL of total 10 µM annealed crRNA-tracrRNA (gRNA) (e.g., 0.5 µM each for 20 guides) with 3 µL 10x CutSmart buffer (NEB) and 0.3 µL 62 µM HiFi Cas9. [For the 1 µM file 10 enrichment experiment, only a total of 1 µM gRNA was used.] Volume was adjusted to 30 µL with nuclease-free water and reaction was incubated at room temperature for 15 min, then kept on ice. High molecular weight (approximately 10 kb) RCA product was dephosphorylated by combining 24 µL (about 5 µg) of RCA pool with 3 µL of 10x CutSmart Buffer (NEB) and 3 µL of Quick CP (NEB) and was incubated at 37 °C for 10 min, then 80 °C for 2 min, then held at room temperature (20 °C). To cleave and dA-tail the RCA sample, the entire dephosphorylated product was gently mixed with 10 µL of the Cas9 RNPs, 1 µL of 10 mM dATP and 1 µL Taq polymerase (NEB). The reactions were incubated at 37 °C for 1 min (12-file and single-file targets) or 15 min (3-file and 20-file targets), then at 72 °C for 5 min, then held at 4 °C or on ice. Adapter mix was prepared in a separate tube by well-mixing 20 µL of Ligation Buffer, 3 µL of nuclease-free water, 10 µL of NEBNext Quick T4 DNA Ligase and 5 µL of AMX adapters. The adapter mix was combined with the cleaved and dA-tailed product for a total volume of 80 µL and incubated at room temperature for 10 min. The ligation yield was purified and concentrated using AMPure XP beads at 0.8 beads:1 sample (ligation yield + 80 µL nuclease-free TE buffer) and eluted in 12 µL Elution Buffer.

Nanopore sequencing

Nanopore sequencing was performed on R9.4.1 flow cells from ONT. The flow cells were primed by loading 800 µL from a mix of 1170 µL of Flush Buffer and 30 µL of Flush Tether into the priming port and waiting 5 min. The remaining 200 µL of the priming mix was loaded directly before the sequencing sample. For Random Access, 25 µL of Sequencing Buffer, 13 µL of resuspended Loading Beads, and 12 µL of the eluted DNA library were combined. For Similarity Search, 37.5 µL Sequencing Buffer, 25.5 µL resuspended Loading Beads, and 12 µL of the eluted DNA library were combined. Sample was loaded dropwise on the SpotON sample port. Sequencing was run at 37 °C for 18–24 h. When not in use, flow cells were stored in C18 buffer (150 mM potassium ferrocyanide, 150 mM potassium ferricyanide, 25 mM potassium phosphate, pH 8.0) at 4 °C.

Random access basecalling and data analysis

Basecalling on sequencing reads was performed using Guppy v3.2.2 (available from ONT) with a quality score cutoff of 9. Reads were then processed using C3POa (https://github.com/rvolden/C3POa), which demultiplexes reads into respective files based on file address and generates a consensus sequence for each concatemeric read. Following C3POa, the splint sequences, primer sequences, and file addresses were trimmed off of each read, leaving the payload. Payload sequences were then decoded to recover the original digital files stored in DNA.

Similarity search datasets

With the exception of the query images, all images were collected from Open Images V4, a dataset of over 9 million URLs for images with Creative Commons licenses. Of these, approximately 1.74 million are hosted by the CVDF and available for download; the rest are raw Flickr URLs and may or may not be available. For the image database used in our experiments, we took all images from the hosted set. For training, we took 1.2 million images from the full set of 9 million that were not used for validation, testing, or experiments.

Similarity search feature extraction

To extract image features, we processed each image with VGG16, a convolutional neural network designed for image classification. The weights were loaded from the publicly available trained model and left unchanged during our processing. We used the activations of FC2 (the second fully-connected layer) as 4096-dimensional feature vectors. Using the same metric as prior work[24], pairs of images with Euclidean distance of 75 or less tend to be consistently similar, so during training, we label these pairs as "similar" and all other pairs as "not similar".

Similarity search sequence encoding

The sequence encoder is a fully connected neural network. The 4096-dimensional FC2 vectors are fed into a 2048-dimensional hidden layer with a rectified linear activation, followed by an output layer with a "one-hot" sequence representation. The output layer has dimensions N by 4, where N denotes the number of nucleotides in the sequence,

which was 20 in our case. In this representation, each sequence position has four channels, one for each base. A straight-through estimator is used to convert each base to a one-hot vector with a hardmax function, while a softmax activation function is applied during backpropagation to estimate the gradients despite the non-differentiable hardmax function used in the forward pass. A DNA sequence can be read off by picking the channel with the maximum activity at each position.

The yield predictor takes a pair of one-hot sequence representations and produces an estimate of the Cas9 activation rate between the first (the target) and second sequence (the query). This score acts as a simulated yield score in a retrieval experiment.

Similarity search encoder training

During each round of encoder training, we draw a batch of pairs of feature vectors from the training set. This batch of pairs is formed by randomly choosing "anchor" images - please note these are not images of anchors. Each anchor image is then paired with a dissimilar image (defined as having a Euclidean distance between the image feature vectors > 75), and a similar image (Euclidean distance between the image feature vectors ≤ 75). This process of using anchor images in this way is known as triplet loss. Due to memory constraints, the training dataset is broken up into 16 batches, one batch is loaded at a time, and training samples are drawn from the currently loaded batch. Throughout training, the batch is periodically changed by randomly selecting a new batch file to load into memory. After selecting random anchors, it is sometimes not possible to find a vector similar to the anchor. If, after a fixed amount of searching for a similar vector one is not found, the anchor vector is duplicated as the similar vector.

The batch of triplets is processed by the encoder, which produces triplets of one-hot sequences. The positive and negative pairings from each triplet of sequences are processed by the yield predictor, which outputs the estimated yield of Cas9 cleavage for the positive and negative pairs in each triplet. The estimated yield is calculated by adapting the wtCas9 cleavage prediction model from prior work[32] to be a differentiable function and thus able to be used by our TensorFlow workflow. Loss for each triplet is calculated using the log of the yield predictor. In order to achieve a large (near to zero) log-activation for the positive, and maintain a small (much less than zero) log-activation for the negative sample, the loss is calculated as the difference: the log-activation of the negative sample minus the log-activation of the positive sample. In order to focus the backpropagation training on harder, not yet learned samples, the contribution of both the negative sample and positive sample are clipped at acceptable levels. Positive log-activation scores greater than −0.5 and negative log-activation scores less than −3.0, are clipped at these values preventing them from contributing to training gradients. The encoder weights are trained using the Adaptive Gradient algorithm (Adagrad).

Similarity search barcodes

Most documents are encoded by a shared DNA sequence (Supplementary Fig. 3). The 457 unique DNA sequences each have a unique file ID of length 20 with a minimum edit distance of 5 from every other ID. Additionally, a 60 nt region is included for each of the 457 sequences. These sequences improve alignment as they are highly orthogonal with a minimum edit distance of 20 nt between any two strands. They also help differentiate between the first and second fragments of a cut strand.

Similarity search DNA library synthesis and amplification

The DNA pool containing 457 unique strands representing our image database was synthesized by Twist Bioscience. The pool was amplified using 2x KAPA Master Mix with 10 µM primers (95 °C for 3 min; then 14 cycles of: 98 °C for 20 s, 62 °C for 15 s, then 72 °C for 30 s; followed by 72 °C for 30 s). DNA was purified using the QIAGEN QIAquick PCR Purification Kit protocol and diluted with nuclease-free water to a final concentration of 5 ng/μL.

Similarity search Cas9 enrichment

Cas9 ribonucleoprotein complexes (RNPs) were prepared by combining 5 µL of 1.5 µM annealed crRNA-tracrRNA (sgRNA) with 1.5 µL of a 5 µM wild-type Cas9 (IDT) solution diluted with 10x CutSmart buffer (NEB). Volume was adjusted to 15 µL with nuclease-free water, and the reaction was incubated at room temperature for 15 min. Cut Master Mix (CMM) was prepared by combining 3.5 µL of dATP (10 mM) with 3.5 µL of Taq Polymerase (NEB) and 8 µL of nuclease-free water. 12 µL of RNPs were combined with 12 µL of CMM and held at room temperature (20 °C). Amplified Twist pool was dephosphorylated by combining 15 µL of 5 ng/μL pool with 3 µL of 10x CutSmart Buffer (NEB), 1.5 µL of Quick CIP (NEB) and 12.5 µL of nuclease-free water and was incubated at 37 °C for 10 min, then at 80 °C for 2 min, then held at room temperature (20 °C). To cleave and dA-tail the sample, the entire dephosphorylated product was gently mixed with 10 µL of the Cas9 RNP+CMM mix. The reactions were incubated at 37 °C for 30 s, 80 °C for 5 min, then held at 4 °C or on ice. Adapter mix was prepared in a separate tube by well-mixing 25 µL of Ligation Buffer, 3 µL of nuclease-free water, 10 µL of NEBNext Quick T4 DNA Ligase and 5 µL of AMX adapters. The adapter mix was combined with the cleaved and dA-tailed product for a total volume of 85 µL and incubated at room temperature for 10 min. The ligation yield was purified and concentrated using 34 µL of AMPure XP beads (0.4X cleanup) and eluted in 15 µL Elution Buffer. Nanopore sequencing and basecalling were performed according to the steps described in the Nanopore Sequencing section of the random access methods.

Similarity search post-sequencing data analysis

Reads were aligned to reference sequences using Guppy on an ONT GridION. A summary of sequenced strand types and filtering can be found in Supplementary Section 6. All sequencing runs were filtered to exclude reads that were unaligned or had low alignment accuracy (≤ 0.9). Reads from the retrieval experiments were further filtered to include only cut fragments from the second half of the sequence. Read counts were normalized for each sequence by calculating their ratios to the total number of reads. Two baseline sequencing runs were used for comparative metrics. The average ratios of each strand within the baseline sequencing pools was taken for the baseline used in the analyses. Due to uneven amplification or our baseline pool, we removed three strands from analysis whose total read counts were found to be more than two standard deviations below the mean in both baseline sequencing results.

Reporting summary

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

Data availability

Data used for the random access portions of this work are available at https://github.com/uwmisl/cas9-random-access. Data for the similarity search results are found at https://github.com/uwmisl/cas9-similarity-search. The image dataset can be found at https://github.com/uwmisl/primo-openimages. The Electrophoresis image is licensed for use under https://creativecommons.org/licenses/by-sa/4.0.

Code availability

Code used to perform analyses on the random access portions of this work are available at https://github.com/uwmisl/cas9-random-access DOI: 10.5281/zenodo.15499096 and https://github.com/uwmisl/cas9-similarity-searchhttps://doi.org/10.5281/zenodo.15499104.

References

1. Organick, L. et al. Probing the physical limits of reliable DNA data retrieval. *Nat. Commun.* **11**, 1–7 (2020).
2. 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).
3. Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. Nucleic acid memory. *Nat. Mater.* **15**, 366–370 (2016).
4. Rutten, M. G. T. A., Vaandrager, F. W., Elemans, J. A. A. W. & Nolte, R. J. M. Encoding information into polymers. *Nat. Rev. Chem.* **2**, 365–381 (2018).
5. Kohll, A. X. et al. Stabilizing synthetic DNA for long-term data storage with Earth alkaline salts. *Chem. Commun.* **56**, 3613–3616 (2020).
6. Organick, L. et al. An empirical comparison of preservation methods for synthetic DNA data storage. *Small Methods* **5**, 2001094 (2021).
7. Yazdi, S. M. H. T., Yuan, Y., Ma, J. & Zhao, H. A rewritable, random-access DNA-based storage system. *Nature Publishing Group* 1–10 https://doi.org/10.1038/srep14138 (2015).
8. Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77–80 (2013).
9. Bornholt, J. et al. Toward a DNA-based archival storage system. *IEEE Micro* **37**, 98–104 (2017).
10. Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. *Science* **337**, 1628 (2012).
11. Organick, L. et al. Random access in large-scale DNA data storage. *Nat. Biotechnol.* **36**, 242–248 (2018).
12. Erlich, Y. & Zielinski, D. DNA fountain enables a robust and efficient storage architecture. *Science* **355**, 950–954 (2017).
13. Yazdi, S. M. H. T., Gabrys, R. & Milenkovic, O. Portable and error-free DNA-based data storage. *Sci. Rep.* **7**, 5011 (2017).
14. Tomek, K. J. et al. Driving the Scalability of DNA-Based Information Storage Systems. *ACS Synth. Biol.* **8**, 1241–1248 (2019).
15. Carr, P. A. & Church, G. M. Genome engineering. *Nat. Biotechnol.* **27**, 1151–1162 (2009).
16. 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).
17. Shendure, J. et al. DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).
18. Winston, C. et al. Combinatorial PCR Method for Efficient, Selective Oligo Retrieval from Complex Oligo Pools. *ACS Synth. Biol.* **11**, 1727–1734 (2022).
19. Takahashi, C. N., Nguyen, B. H., Strauss, K. & Ceze, L. Demonstration of end-to-end automation of DNA data storage. *Sci. Rep.* **9**, 4998 (2019).
20. Antkowiak, P. L. et al. Integrating DNA encapsulates and digital microfluidics for automated data storage in DNA. *Small* **18**, 2107381 (2022).
21. Xu, Q., Schlabach, M. R., Hannon, G. J. & Elledge, S. J. Design of 240,000 orthogonal 25mer DNA barcode probes. *Proc. Natl Acad. Sci. USA* **106**, 2289–2294 (2009).
22. Banal, J. L. et al. Random access DNA memory using boolean search in an archival file storage system. *Nat. Mater.* **20**, 1272–1280 (2021).
23. Bögels, B. W. A. et al. DNA storage in thermoresponsive microcapsules for repeated random multiplexed data access. *Nat. Nanotechnol.* **18**, 912–921 (2023).
24. Bee, C. et al. Molecular-level similarity search brings computing to DNA data storage. *Nat. Commun.* **12**, 4764 (2021).
25. Sadremomtaz, A. et al. Digital data storage on DNA tape using CRISPR base editors. *Nat. Commun.* **14**, 6472 (2023).
26. Hou, Z. et al. "Cell Disk" DNA storage system capable of random reading and rewriting. *Adv. Sci.* **11**, 2305921 (2024).
27. Zhang, J., Hou, C. & Liu, C. CRISPR-powered quantitative keyword search engine in DNA data storage. *Nat. Commun.* **15**, 2376 (2024).
28. Hsu, P. D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
29. Pattanayak, V. et al. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).
30. Chen, J. S. et al. Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).
31. Fu, B. X., St Onge, R. P., Fire, A. Z. & Smith, J. D. Distinct patterns of Cas9 mismatch tolerance in vitro and in vivo. *Nucleic Acids Res.* **44**, 5365–5377 (2016).
32. Jones, S. K. et al. Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *Nat. Biotechnol.* **39**, 84–93 (2021).
33. Gilpatrick, T. et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.* **38**, 433–438 (2020).
34. K. O. on & 2020. #MemoriesInDNA portrait project blends DNA technology and art to memorialize pioneering scientist Rosalind Franklin. Allen School News https://news.cs.washington.edu/2020/02/24/memoriesindna-portrait-project-blends-dna-technology-and-art-to-memorialize-pioneering-scientist-rosalind-franklin/.
35. Volden, R. et al. Improving nanopore read accuracy with the r2c2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl Acad. Sci. USA* **115**, 9726–9731 (2018).
36. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
37. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. preprint at *arXiv* https://doi.org/10.48550/arXiv.1409.1556 (2015).
38. Tan, M. & Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*. 6105–6114 (PMLR, 2019).
39. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90.

Acknowledgements

We are greatful to Martin Dahinden, Michelle Beauregard, and Erik Dahinden for their generosity in allowing us to use the image of Bigfoot. Funding was provided by National Science Foundation grant 2212306 (L.C., J.N.) and the National Science Foundation's Graduate Research Fellowship DGE-2140004 (C.I.).

Author contributions

N.C and K.Z. and C.A. conceived of, designed, performed and analyzed random access experiments and wrote the manuscript. L.O. and C.I conceived of, designed, performed and analyzed similarity search experiments and simulations and wrote the manuscript. J.M. designed, performed and analyzed similarity search simulations and built the similarity search code base. C.B. and J.D. built the similarity search code base. D.W., G.R. and S.J. performed similarity search lab experiments. K.S. and L.C. supervised the work. J.N. directed and supervised the work and wrote the manuscript.

Competing interests

K.S. is currently employed by Microsoft. J.N. is a consultant for Oxford Nanopore Technologies. All other authors declare no competing interests.'

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-61264-5.

Correspondence and requests for materials should be addressed to Jeff Nivala.

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