Nucleic Acid Databases and Molecular-Scale Computing

Xin Song[1,2,3*], John Reif[1,3]

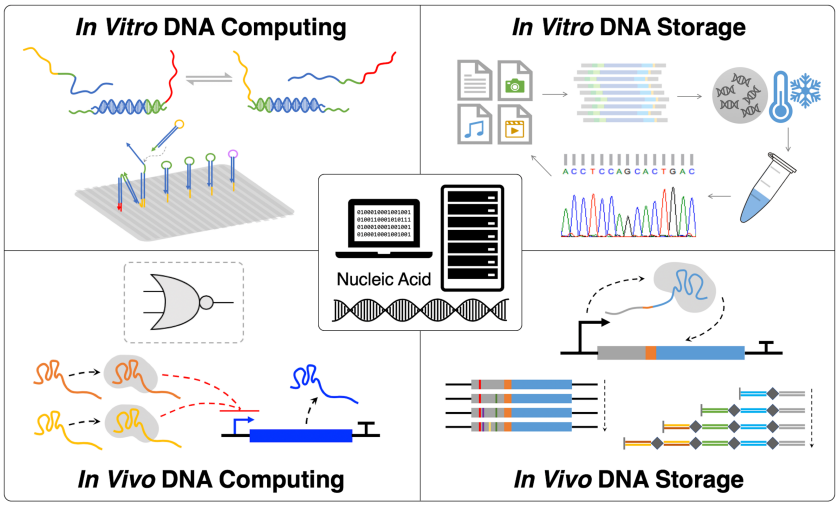
[1]Department of Electrical and Computer Engineering
[2]Department of Biomedical Engineering
[3]Department of Computer Science

Duke University, Durham, NC 27708, USA

[*]E-mail: xin.song@duke.edu

Abstract

DNA outperforms most conventional storage media in terms of information retention time, physical density, and volumetric coding capacity. Advances in synthesis and sequencing technologies have enabled implementations of large synthetic DNA databases with impressive storage capacity and reliable data recovery. Several robust DNA storage architectures featuring random access, error correction, and content-rewritability have been constructed with potential for scalability and cost reduction. We survey these recent achievements and discuss alternative routes for overcoming the hurdles of engineering practical DNA storage systems. We also review recent exciting work on *in vivo* DNA memory including intracellular recorders constructed by programmable genome editing tools. Besides information storage, DNA could serve as a versatile molecular computing substrate. We highlight several state-of-the-art DNA computing techniques such as strand displacement, localized hybridization chain reactions, and enzymatic reaction networks. We summarize how these simple primitives have facilitated rational designs and implementations of *in vitro* DNA reaction networks that emulate digital/analog circuits, artificial neural networks, or non-linear dynamic systems. We envision these modular primitives could be strategically adapted for sophisticated database operations and massively parallel computations on DNA databases. We also highlight *in vivo* DNA computing modules such as CRISPR logic gates for building scalable genetic circuits in living cells. To conclude, we discuss various implications and challenges of DNA-based storage and computing, and we particularly encourage innovative work on bridging these two areas of research to further explore molecular parallelism and near-data processing. Such integrated molecular systems could lead to far-reaching applications in biocomputing, security, and medicine.



Keywords: DNA nanotechnology, synthetic biology, digital data storage, molecular computer, strand displacement, hairpin hybridization, localized reaction network, cellular memory, genome editing, CRISPR

Early humankind harnessed materials such as rock, bone, and clay to record information before paper was invented around 2000 years ago. Modern storage systems[1] rely on magnetic, electronic, or optical media for data preservation (Table 1). However, conventional digital storage media generally have limited lifespans due to material degradation and technology obsolescence.[2] With rising demand for big data storage, modern datacenters may become unsustainable due to infrastructure cost and power consumption. Our society now faces pressing need for alternative storage media that are durable, scalable, and economical. DNA has attracted immense interest as a promising non-volatile storage medium for its long-term durability, enormous storage capacity, and remarkable volumetric density.[3] Recent advances in nucleic acid technologies have enabled researchers to encode nonbiological information in synthetic oligonucleotides (oligos) or genomic DNA of living organisms. Early synthetic DNA databases[4,5] relied on coding redundancy and deep sequencing to mitigate errors and recover data from large pools of oligos. Recent implementations[6–9] leveraged compact encoding to improve coding capacity[10] and introduced error-detection/correction schemes for error-free data reconstruction. Polymerase chain reaction (PCR) and high-throughput sequencing (HTS) have enabled robust random access in large-scale DNA databases.[9,11–13] With targeted DNA editing tools, scientists have engineered rewritable DNA databases both *in vitro*[11] and *in vivo*.[14] DNA-based *in vivo* memory can record dynamic temporal events into the genomic DNA of living cells, enabling exciting applications ranging from molecular biosensing to cell lineage tracing.[15] Synthetic DNA has also been exploited as a versatile substrate for molecular computing since Adleman's pioneering work in 1994.[16] The highly specific Watson-Crick base paring allows researchers to systematically program DNA hybridization reactions and design networks of interacting DNA molecules to compute in aqueous solutions. *In vitro* DNA computing systems typically leverage hybridization reactions, strand displacement, or reactive hairpin cascades to emulate complex systems such as digital/analog circuits[17] and artificial neural networks.[18] Techniques such as spatially localized hybridization reactions[19] have been harnessed to reliably speed up DNA computation. Furthermore, enzymatic reactions could be utilized to construct *in vitro* DNA reaction networks with non-linear dynamic behaviors.[20] While challenges remain, we believe the time is ripe for bridging large synthetic DNA databases with DNA computing to support massively parallel database operations[21] and unveil the power of molecular near-data processing.[22] On the other hand, *in vivo* computing systems can be engineered as gene circuits for various applications such as cell programming,[23] and interfacing with DNA-based molecular recorders may create exciting opportunities for more sophisticated intracellular computations and smart therapeutics.[24,25]

Enabling Technologies and Techniques

Long-Term Preservation of DNA

DNA could store information for extremely long term.[26,27] However, variations in storage conditions (*e.g.* humidity, oxidation, temperature, and light) can affect the stability and integrity of DNA molecules. In aqueous solutions, DNA molecules are prone to hydrolytic cleavage, depurination, depyrimidination, and deamination.[28] Techniques such as dehydration and encapsulation could minimize the humidity effect on degradation. Oxidative damage can cause DNA mutations. Methods for reducing oxidation include adding chelators or antioxidants in storage media[28] or encapsulating DNA with inorganic materials.[29] Temperature also plays an important role in DNA preservation. Low temperature restricts the molecular mobility of DNA to slow down degradation.[28] However, maintaining an extremely cold environment could be costly and challenging to scale, whereas techniques such as dehydration and encapsulation are more practical for inexpensive long-term DNA preservation at large scales.

DNA Sequence Design, Synthesis, and Sequencing

The field of DNA computing has benefited from publicly available software packages[30–32] that design, analyze, and simulate systems of interacting oligos based on the thermodynamic properties of nucleic acids.

Modular techniques such as domain-based sequence designs[33] provide simple and systematic frameworks for constructing complex chemical reaction networks (CRNs) with synthetic oligos. Innovations in instrumentation and chemistry have enabled large-scale DNA synthesis and sequencing with lower cost and higher accuracy. Oligos designed and verified *in silico* can be synthesized on column-based or array-based synthesizers with varying tradeoffs in synthesis scale, error rate, and cost. Nowadays, microarray-based technology can synthesize large pools of distinct oligos simultaneously. DNA *de novo* synthesis typically involves the conventional solid-phase phosphoramidite chemistry or light-activated chemistries.[34] Recent milestones in enzymatic synthesis[35] may soon offer fast and accurate synthesis with tremendous cost reduction. DNA sequencing has been more affordable than synthesis and has generated an enormous amount of data to help scientists decipher the genetic blueprints of life and engineer biology.[34] HTS platforms such as Illumina sequencers offer superior sequencing speed and throughput with low error rate, and single-molecule sequencers such as the portable Oxford Nanopore MinION could read long sequences in real-time. Recent advances in DNA synthesis and sequencing technologies have been extensively reviewed.[34,36] Large-scale DNA storage becomes more practical as sequencing and synthesis continuously improve in throughput, accuracy, and affordability.

Detection, Amplification, and Quantification of DNA

As one of the most indispensable tools in biological sciences, PCR enables sensitive detection and rapid amplification of specific oligos from large and complex pools of DNA.[37] Basic components of PCR include a template, nucleotides, primers, and thermostable DNA polymerase. PCR reactions rely on well-designed primers to amplify the target sequence with high specificity and yield. High-fidelity polymerases offer proofreading mechanisms to reduce bias and improve the precision, sensitivity, and speed of PCR. The three steps of PCR (DNA denaturation, annealing, and extension) can be automated to run multiple cycles on a thermocycler to exponentially amplify trace amount of template DNA. The resulting PCR product can be quantified by quantitative PCR (qPCR), visualized by gel electrophoresis, and analyzed by sequencing.[37] In DNA-based storage systems, PCR amplification could be leveraged to add information redundancy and cheaply replicate the data stored in DNA. Owing to its high specificity and scalability, PCR amplification has also been adapted to enable random access in recent synthetic DNA databases.[9,11–13] Particularly, nested PCR allows for efficient data addressing and retrieval in hierarchical and multidimensional DNA storage.[38]

Targeted Editing of DNA

Besides the long-term archival capacity, DNA-based storage could also support content rewritability *via* DNA editing to selectively manipulate storage sequences. Conventionally, targeted mutagenesis involves creating a double-stranded break (DSB) near the target site and relying on intracellular repair pathways such as homologous recombination (HR) or non-homologous end-joining (NHEJ) to introduce base substitutions, insertions or deletions (indels).[39] Precision editing tools could generate uniquely targeted breaks by sequence recognition. Early development of recombinant nucleases such as zinc finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) leveraged sequence-specific DNA-binding proteins fused with nuclease domain to achieve programmable editing on virtually any sequence.[40] However, gene editing with site-specific nucleases was labor-intensive and time-consuming because different chimeric nucleases must be engineered for new sequences. Since scientists discovered CRISPR (clustered regularly interspaced short palindromic repeat) as a highly adaptive bacterial defense mechanism against viral infection,[41] the endogenous CRISPR/Cas system has been repurposed to implement gene editing tools with tremendous programmability and ease of use. If the target sequence resides in proximity to a protospacer-adjacent motif (PAM), one could easily design a complementary single guide RNA (sgRNA) to create Cas-mediated DSB at the target site. To expand the targeting scope, variants and orthologs of the wild-type Cas protein have been adapted to provide novel PAM sequences. Inspired by CRISPR's arrayed-spacer architecture, multiplexed CRISPR/Cas could target and edit different sites

simultaneously.[42] Alternatively, pooled sgRNA libraries can be used for large-scale editing in genome-wide screens.[43] Moreover, scientists have achieved single-base editing using CRISPR with fusion enzymes engineered from catalytically-dead Cas (dCas) proteins.[44,45] Advances and applications of CRISPR-based gene editing tools have been surveyed in excellent reviews.[46,47] Notably, researchers have leveraged mechanisms such as directed protospacer acquisition[48–50] and self-targeting CRISPR/Cas[51,52] to implement scalable cellular memories, enabling exciting applications such as cell lineage tracing and brain mapping. Various CRISPR-based DNA storage[14] and computing[53] systems have been demonstrated *in vivo*.

DNA Computation and Databases

In Vitro DNA Computing

Interactions between single-stranded DNA (ssDNA) molecules are highly predictable and programmable owing to the specificity of Watson-Crick base pairs. DNA-based computing systems typically use synthetic ssDNA molecules as inputs and outputs and compute *via* networks of hybridization reactions such as strand displacement cascades.[54] Domain-based sequence abstractions[33] simplify the design of complex and scalable architectures. Tools such as Visual DSD[55] can prototype, model, and analyze DNA-based reaction networks capable of solving sophisticated mathematical functions[56] or implementing digital[57] or analog[58] circuits. In particular, the toehold-mediated strand displacement[59] (TMSD) mechanism has led to both diffusion-based and spatially-localized DNA computing systems. In TMSD (Figure 1a), a signal strand binds to the exposed toehold of a gate complex, undergoes a random walk process (branch migration), and displaces the initially gate-bound strand as an output or input to downstream circuitries. Separation of a dye-quencher pair in the gate complex results in measurable fluorescence for output quantification. The reaction kinetics of reversible TMSD can be modulated using techniques such as toehold sequestering[57] and toehold exchange,[60] and the reaction rates can be tuned many orders of magnitude by varying the length or sequence composition of toeholds.[61] These simple and versatile primitives make TMSD a powerful programming language for designing complex and scalable DNA computing architectures. For instance, DNA seesaw logic circuits[17] can be constructed with three basic reactions – seesawing, thresholding, and reporting (Figure 1b). Toehold exchange participates in not only the reversible seesawing reactions but also the entropy-driven catalytic reactions for signal amplification. Toehold sequestering implements irreversible strand displacement for signal thresholding and reporting. Thresholding defines the logic function of gate motifs and maintains the digital logic abstraction of signals. Multilayer digital circuits rely on signal thresholding and amplification to achieve robust signal restoration. *In vitro* demonstrations of a square root calculator[62] and Hopfield associative memory[63] illustrated the flexibility and robustness of complex multilayer seesaw circuits. Various strategies have been proposed to construct renewable,[64] time-responsive,[65] or arbitrary probabilistic switching circuits[66] based on the seesaw motif. A recent work extended the motif to implement winner-take-all neural networks capable of classifying sophisticated and noisy patterns such as handwritten digits.[18]

The reaction kinetics of free-floating strands limits the speed of DNA computing in solution. Attempts to improve speed by raising temperature or increasing DNA concentrations may introduce higher leaks (spurious interactions between DNA strands) and degrade circuit performance. To reliably speed up computation in large circuits, viable strategies include the use of single-rate TMSD reactions[66] and localized hybridization reactions.[67–70] Recently, "DNA domino"[19] was proposed as a robust localized architecture. The authors constructed circuits with reactive hairpins co-localized on DNA origami surface *via* extended staple strands (Figure 1c). Each hairpin harbored an exposed toehold, a double-stranded stem, and a sequestered toehold in the loop. During hairpin hybridization chain reactions, an input strand opens the first hairpin, which exposes its sequestered toehold to bind and open the second hairpin. The cycle goes on until the output hairpin is reached. The authors arranged the hairpins in specific patterns and adapted standard design strategies of TMSD reactions to construct various circuit modules including multi-input logic gates

and signal transmission lines. Such spatial organization favored reactions between proximal hairpins, effectively speeding up the computing and mitigating interferences between far-apart circuit elements.[19] Notably, their architecture leveraged free-floating fuel hairpins to transmit signals between neighboring hairpins on DNA origami, resulting in significant leak reduction to facilitate rapid and reliable computation even at low concentrations. Moreover, the spatially localized hairpins allowed for the reuse of identical sequences without cross-talk. For example, standard transmission lines were constructed with identical intermediate hairpins, and a small set of orthogonal hairpins were reused to build signal crossover junctions for circuit layout optimization.[19] This offered excellent modularity and scalability compared with diffusion-based systems which typically require stringent mutual orthogonality of circuit components. Others have demonstrated alternative substrate materials to localize and accelerate DNA-based computing.[71]

Besides sequence recognition, other properties of DNA (*e.g.* melting temperature[72] and secondary structures[73]) or enzymes (*e.g.* polymerase[74] and DNAzyme[75]) may be harnessed to explore molecular computing. For example, specific ssDNA sequences may fold into conformations that exhibit catalytic activities to cleave target nucleic acid substrates upon binding. These catalytic DNAs (*i.e.* DNAzymes) could be utilized to regulate DNA circuits by degrading or releasing specific DNA reactants in the network. Libraries of DNAzymes have been designed to construct logic gates and multilayer circuits.[76] Using simple enzymatic reactions, researchers have also assembled *in vitro* DNA reaction networks with non-linear dynamic behaviors. Remarkably, Montagne *et al.*[20] proposed a generic framework named the "PEN DNA toolbox" for rational designs and *in vitro* constructions of dynamic DNA circuits. Standard DNA biochemistry involving polymerase, exonuclease, and nickase was applied to implement basic modules for DNA species activation, inhibition, and destruction (Figure 1d). Specifically, the activation module utilized polymerase and nickase to continuously generate ssDNA output from a DNA template *via* autocatalytic amplification. The inhibition module introduced strands that compete with the input for template binding but have 3'-mismatches to avoid priming the output production. The destruction module leveraged exonuclease to annihilate unprotected ssDNA into inactive mononucleotides. These modules could be cascaded to form *in vitro* DNA reaction networks with positive/negative loops and delays, leading to interesting dynamics such as sustained oscillations. Moreover, the simplicity of modules allowed for quantitative analysis and accurate prediction of system dynamics using mathematical models, which provided insights on parameter optimizations (*e.g.* sequence designs and template/enzyme concentrations) to fine-tune the behaviors of the designed system. *In vitro* DNA reaction networks with dynamic behaviors could be useful tools for implementing DNA database instructions that require precisely orchestrated sequence of operations.

In Vitro DNA Databases

Since Baum[77] envisioned vast associative memories built with synthetic DNA, several DNA storage schemes have been designed and experimentally demonstrated.[78,79] Here we review recent implementations of large DNA databases (Figure 2) highlighting key metrics such as storage capacity, encoding density, error correction, random access, and content rewritability. Details are summarized in Table 2.

In 2012, Church *et al.*[4] used synthetic oligos to encode data containing a book, 11 images, and a computer program. They split data into chunks to avoid long sequences and encoded only 1 bit per base to mitigate undesired sequence patterns. An address was incorporated in each oligo to order sequence assembly. However, this approach could not support random access as the entire pool must be sequenced and decoded to retrieve any file. After amplification and consensus alignment, the authors recovered the original 5.27 million bits with 10 bit-errors. Goldman *et al.*[5] stored various files types (ASCII text, PDF file, JPEG picture, and MP3 audio) in their DNA storage with error correction. Using a customized Huffman code, they translated each byte to a series of base-3 digits, which were converted to nucleotides *via* a simple rotating code to avoid homopolymers. The resulting long string was split into overlapping segments to provide

fourfold coding redundancy. Alternating segments were reverse-complemented to reduce systematic errors, and majority voting was used to correct errors during sequence reassembly. Each data-encoding oligo contained a two-part address to identify files and intrafile locations and a parity-check to detect errors. The authors reconstruct all files accurately despite two 25-nt regions that needed manual intervention to recover. The regions contained repeats forming self-reverse-complementary patterns that failed synthesis, and the authors suggested input randomization to avoid such repeats during encoding. To illustrate the storage robustness and cost efficiency, they subsampled the reads more than 10-folds to simulate low sequencing coverage and still reconstructed the data perfectly. To store different data types with optimized coding density, Bornholt *et al.*[12] leveraged separated oligo pools and tunable data redundancy. In their key-value architecture, each key determined the storage pool for a file's coding strands and the assignment of primers, enabling efficient PCR-based random access in individual pools. The authors introduced block-level redundancy *via* XOR operations between payloads and encoded the results into new strands. Coding redundancy could be fine-tuned based on the data block importance. Compared to the Goldman[5] coding, this work encoded information twice as dense and achieved complete data recovery with minor intervention.

Yazdi *et al.*[11] demonstrated a DNA database supporting both error-free random access and content rewriting. They designed mutually uncorrelated primers and applied prefix-synchronized codes on data blocks to minimize sequence cross-hybridization. Content editing was achieved using gBlocks or OE-PCR. However, the dictionary-based encoding was limited to storing text. In another work, Yazdi *et al.*[9] improved the coding strategy and recovered two compressed images with an error-prone nanopore sequencer. The constrained coding reduced homopolymers and balanced the GC-contents of codewords. The use of long codewords (1000 bp) enabled highly efficient coding, and mathematically constructed addresses supported robust random access. To tolerate sequencing errors, they devised an integrated pipeline of consensus alignment to estimate codewords. After identifying high-quality reads, several multiple-sequence-alignment algorithms were used to generate different consensus sequences. An aggregate consensus was then generated by majority voting with GC-balancing constraint and further improved by BWA alignment and error correction. This alignment strategy reduced most insertion and substitution errors to deletion errors, which were easily corrected by homopolymer checks. Despite the high error rate of MinION sequencer, this work demonstrated error-free data recovery with a coding density of $1.1*10^{23}$ bytes/gram.

Grass *et al.*[6] investigated long-term chemical preservation of DNA storage with integrated error-correcting codes. They applied concatenated RS codes to add redundancies accounting for both single-base errors and loss of complete sequences. Specifically, their inner code could correct more than 3 arbitrary base errors per sequence, and the outer code could further correct about 8.5% sequence errors or handle 17% complete sequence losses. The coding scheme offered robust error tolerance but no random access. For long-term storage, the oligos were encapsulated in silica particles to protect DNA against both humidity and oxidation. After a week of accelerated aging experiments to simulate DNA decay over 4 half-lives, the authors released the oligos *via* fluoride etching and completely recovered the stored data. By extrapolation, they estimated that their DNA storage could last for 2000 years around 10°C or 2 million years in the Global Seed Vault. Blawat *et al.*[7] demonstrated robust storage and error-free recovery of a 22MB compressed movie, achieving a remarkable leap in DNA storage capacity. Based on their systematic analysis of experimental data, the proposed forward-error-correcting schemes were tailored to mitigate all types of errors arising from DNA synthesis, PCR amplification, and sequencing. Each data byte mapped to a 5-nucleotide chunk (DNA symbol) during encoding. Error propagation caused by base substitutions was minimized by strategic mapping of two-bit tuples to nucleotides at different positions in symbols. Using symbols from alternating groups to code binary bits enabled detection of indel errors and helped eliminate self-reverse-complementary sequences. Each oligo contained three parts: address, data payload, and error-detecting code. A strong BCH code was used for address protection, an RS block code for coding redundancy of consecutive payloads, and a cyclic redundancy check for parity checking in each oligo. Such coding scheme achieved a remarkably small residual error probability for use in practical DNA storage systems.

Erlich and Zielinski[8] presented an elegant DNA storage strategy based on fountain codes. They stored 2.14MB of compressed data (including an operating system, malware, movie, and PDF/text/image files) at a remarkable density equavalent to 86% of information capacity in nucleotides. They divided the binary data into non-overlapping segments and each time selected a few using a random number generated from special distributions. Each set of selected segments was encoded as a "droplet" containing (i) the bitwise-XOR result of segments as payload, (ii) the pseudorandom-number-generator seed for segments identification, and (iii) an RS code for error correction. The entire droplet formation process was described by a Luby transform, which was repeated to generate a large pool of binary droplets. A direct mapping from {00,01,10,11} to {A,C,G,T} converted the droplets to oligos. During this process, problematic oligos with homopolymers or undesired GC-content were discarded and the transform was repeated until valid oligos were found. The nature of fountain codes allowed data reconstruction by collecting enough droplets to reverse the Luby transform. This coding scheme supported highly tunable redundancy by generating different numbers of oligos without complicating the algorithm design. Decoding was highly robust as one could simply discard erroneous oligos and analyze the high-quality reads. The authors further tested their storage against serial PCR amplifications and dilutions, demonstrating nearly unlimited data retrievals and error-free recovery at a maximal physical density of $2.15*10^{17}$ bytes/gram.

More recently, Organick *et al.*[13] scaled up DNA storage capacity with robust random access. They encoded over 200MB of compressed data (35 files of various sizes/types) and achieved error-free decoding with merely 5x sequencing coverage. They designed a pipeline to evolve and optimize a large number of orthogonal primers, which were scored and screened in terms of homopolymers, self-complementarity, GC content, *etc.* To encode data, they pseudorandomized and segmented the binary bits, applyed an RS outer code to introduce redundancy and an inner code to convert bits to nucleotides, and then assigned each oligo with primers to enable PCR-based random access. Because of the low coverage, their decoder was designed to maximally utilize available reads including noisy reads and iteratively cluster them based on similarity. To estimate the original sequence, a consensus was generated from each cluster by trace reconstruction. The authors then reverted the inner/outer codes and randomization to reconstruct the data. To further test the error tolerance, they assembled the oligos of two files into long sequences, sequenced them using error-prone MinION, and successfully recovered the data despite low read coverage. Researchers from the same group has reported successful storage of over 400 MB digital data in DNA[22] at the time of writing.

In Vivo DNA Memory and Computing

Genome editing has enabled scientists to engineer DNA-based storage systems in living organisms. Early work on cellular memories leveraged inversion recombination[80] to record binary states into genomic DNA. For instance, Bonnet *et al.*[81] demonstrated a rewritable digital register by repeatedly inverting and restoring a DNA segment in *E. coli* genome. To achieve controlled switching between two states, the authors leveraged bacteriophage integrase with an excisionase cofactor to modulate the directionality of recombination reactions. Specifically, expression of the integrase alone would flip the DNA segment with pre-engineered recognition sites, and the co-expression of integrase and excisionase would revert the segment back to its original orientation. Such DNA-based latch storage supported repeated switching *in vivo* and maintained state over 100 cell generations. Siuti *et al.*[24] went a step further and built integrated logic/memory systems in *E. coli* cells by combining various DNA-inversion-based functional modules. Others engineered larger memory arrays[82] and cellular state machines[83] *via* specific concatenation or overlapping of recombinase recognition sites. However, recombinase-based systems must rely on highly orthogonal enzymes and long stretches of DNA to encode single bits, which limited the memory scalability and underutilized the information capacity of DNA. Farzadfard and Lu[84] devised a system termed "SCRIBE" to implement scalable analog memory *in vivo*. Their system leveraged retron-mediated expression of ssDNA templates with co-expressed recombinase to induce targeted mutations on genomic DNA upon regulatory inputs such as transcriptional signals or light. Complementary SCRIBE modules enabled

repeated rewriting, and different template ssDNAs were designed for multiplexed recording at independent loci, offering programmability and scalability. Analog information (*e.g.* magnitudes and durations of transient signals) was recorded in the form of accumulated mutations distributed across cell populations.

To engineer robust cellular memory, scientists have also garnered insights from immunological mechanisms of living cells, and in particular the CRISPR. Shipman *et al.*[48] exploited the directed spacer acquisition in CRISPR arrays (Figure 3a) to record temporal information in bacterial genome, achieving a dramatic improvement in storage capacity. Arbitrary information could be coded into synthetic oligos of defined length. After electroporation into cells over-expressing the Cas1-Cas2 protein complex, these synthetic oligos functioned as protospacers and integrated into the expanding CRISPR array. Because new spacers always appended to the array's leading end, the ordering of integrated spacers could reflect the temporal history of acquisition events over time. The authors noticed adding a 5'-PAM in protospacer could enhance the acquisition frequency and determine the spacer orientation during integration. Hence, their proposed system could record in multiple modalities (*i.e.* the sequence content, temporal ordering, and orientation of integrated spacers). Because spacer acquisitions occurred stochastically at single cell level, memory readout relied on analysis of CRISPR arrays from a population of cells. Focusing on storage scalability, Shipman *et al.* optimized their protospacer design and data reconstruction strategy in a follow-up study.[49] They stored a short GIF movie into bacterial genomes by encoding the frame pixels as serially-electroporated synthetic protospacers. Sequencing of the CRISPR arrays recovered the stored data with high accuracy. Sheth *et al.*[50] developed the "TRACE" framework to record biological signals in CRISPR arrays by integrating intracellularly generated DNA segments. The expression of their DNA spacer (trigger DNA) was modulated by the temporal characteristics of intracellular signals, effectively transforming the temporal biological signals into the abundance changes of trigger DNA. To record time intervals representing the absence of signals, reference spacers were acquired into the genomic CRISPR array at a background rate. The authors were able to reconstruct the dynamic temporal history of cellular signals by analyzing the frequencies and the ordering of acquired spacers in CRISPR arrays among a cell population. They also achieved multiplexed temporal recording with high accuracy using barcoded CRISPR arrays.

Directed spacer acquisition is not the only way to implement CRISPR-based storage. Perli *et al.*[51] leveraged self-targeting CRISPR/Cas (Figure 3b) to design programmable and multiplexed memory architectures for continuous longitudinal recording in human cells. The authors introduced a PAM sequence at the sgRNA-encoding locus to repeatedly direct the Cas nuclease activity toward its own sgRNA-encoding region on genome. This led to accumulated mutations at the self-targeting sgRNA (stgRNA) locus. By coupling the expression of stgRNA or Cas9 to different molecular inducers, various analog events such as the durations and intensities of cellular activities could be continuously recorded *in vivo*. Multiplexed recording could be achieved on independent stgRNA loci simultaneously. The authors reconstructed the recording by analyzing the evolution pattern of the sgRNA locus in a cell population. Despite the novel design, several limitations were pointed out. For example, repeated self-targeting events could shorten the stgRNAs over time and result in compromised targeting specificity or loss of PAM. Long-term recording had to rely on longer stgRNAs, which may complicate sequence design and not scale well. To improve storage capacity and data interpretation, the authors suggested to use techniques such as CRISPR base-editing to introduce more defined mutagenesis on the stgRNA. Frieda *et al.*[85] proposed a system termed "MEMOIR" to record cellular states on barcoded scratchpads *via* CRISPR/Cas-mediated mutagenesis. The authors were able to track cell lineages and dynamic cellular event history by analyzing the collapse patterns of scratchpads *via in situ* single-cell readout without disruptive sequencing. Tang and Liu[86] developed a framework called "CAMERA" to implement rewritable analog recorders *via* CRISPR-mediated manipulation of multicopy plasmids. In their first strategy, analog information such as signal amplitude or duration was translated into the change of copy-number-ratio between two mutually exclusive plasmids. The two recording plasmids were designed with nearly identical sequences such that they could stably coexist in cells while allowing the stimuli-induced CRISPR/Cas activity to selectively cleave only one of them. Such plasmid compensation system reliably recorded multiple analog signals and supported repeated cycles of

erasing/rewriting. In their second strategy, the authors engineered writing plasmids harboring CRIPSR base-editors to record signals as base mutations on the recording plasmids. Although the use of orthogonal inducible regulators limited the recording multiplexity, the CAMERA framework enabled sensitive and reliable recording with small cell populations (10 to 100 cells) owing to the large number of recording plasmids within each cell.

The programmability of sgRNAs also allows simple *in vivo* constructions of CRISPR-based logic gates and multilayer genetic circuits *via* transcriptional regulations of synthetic promoters as inputs and outputs. For example, input promoters that respond to specific cellular cues could be designed to drive the transcription of different sgRNAs, which then guide dCas9 to selectively repress the expressions of downstream sgRNAs or reporter genes (Figure 3c). These sgRNA-promoter pairs could be cascaded in various configurations to build multi-input logic gates[87] and arithmetic operators such as half adders[88] in living cells. Researchers have also utilized CRISPR base-editors to integrate cellular memory with logic operations,[89] demonstrating a promising step toward engineering more sophisticated and intelligent biocomputers *in vivo*.

DNA Steganography and Cryptography

Because of its excellent storage capacity and compact physical volume, DNA is suitable for applications involving data encryption and information hiding. In 1999, Clelland *et al.*[90] demonstrated DNA-based steganography by encoding secret messages into primer-flanked oligos and mixing them with sonicated human DNA. The genomic DNA fragments provided a complex and noisy background to conceal the synthetic oligos. After pipetting the DNA mixture onto a filter paper, microdots were excised and mailed. Knowing the primer sequences and the encryption key, the intended recipient could retrive the secret oligos by PCR amplification and sequencing and subsequently apply the key to decipher the reads. DNA steganography schemes proposed by Leier *et al.*[91] leveraged PCR and gel electrophoresis to achieve fast decryption and readout without sequencing. The authors designed short dsDNA blocks to represent binary bits and ligated them into a "binary strand" encoding the secret message. The strand was then attached with a key and consealed in a large pool of dummy strands with the same length and similar binary structures to ensure effective hiding. Decryption involved two separate PCR amplifications. The first PCR used forward and reverse primers to target the key and bit-0 block, producing amplicons revealing the position of every bit-0 in the secret on a gel image. The second PCR used primers targeting the key and bit-1 block to reveal the bit-1 positions in the secret. The combined gel images led to the recovery of complete secret bit pattern. Alternatively, if dummy strands shared an identical key with the secret message strand, PCR amplification would result in an encrypted pool and the gel image would show mixed bit patterns from all strands. Given a copy of the dummy pool as decryption key, the intended recipient could graphically subtract the dummy pool readout from the encrypted pool readout to obtain the secret bit pattern. Halvorsen *et al.*[92] demonstrated a convenient DNA cryptography using only gel electrophoresis for readout. In their work, a specific ssDNA served as the decryption key to initiatite self-assembly of DNA nanostructures from a seemingly random pool of strands. Secret binary bits were encoded as the conformational changes of two-state nanostructures, which could be as simlpe as self-assembled dsDNAs. Using DNA strands of different lengths to distinguish bit positions in the secret, the authors achieved decryption of 8 bits per gel lane. To securely transmit longer messages, one could use high-resolution gels and efficient coding schemes to increase the capacity of DNA steganography and cryptography.[92] In another work, Gehani *et al.*[93] analyzed various strategies to enhance the security of DNA steganography and constructed DNA one-time-pads theoretically unbreakable by cryptanalysis. DNA cryptography has also been applied to molecular authentication and barcoding.[94]

Conclusions and Outlook

Compared to conventional storage media, DNA offers exceptional volumetric density, longevity, and power efficiency for long-term data preservation with minimal environmental impacts. Information stored in DNA

can be easily replicated and processed with massive parallelism without relying on intricate circuit wiring and stringent spatial organization as required in conventional storage architectures. These features make DNA a compelling medium for building extremely dense, durable, and versatile storage at the molecular level. However, implementing practical storage systems with DNA faces several limitations and challenges. For example, DNA-based storage is unlikely to outperform electronics in terms of read/write speed. Even with full automation, latencies inevitably accrue during synthesis, sequencing, wet lab sample preparations, *etc.* As a result, DNA-based storage is by far mostly considered for data archival purposes. However, building DNA archival storage at a profitable scale is hampered by the standard chemical synthesis, which is slow, unreliable, expensive, and has barely improved over the past 40 years.[22] To address the synthesis bottleneck, several routes can be explored. First, recent investigations have shown promising uses of a specialized polymerase called the terminal deoxynucleotidyl transferase (TdT) to reliably synthesize custom DNA without templates.[95] With ongoing research and commercialization efforts,[35] such enzymatic approach may revolutionize DNA synthesis to offer immensely larger throughput, higher fidelity, and cost reduction by orders of magnitude. Although it is possible to engineer molecular storage using molecules that are simpler and cheaper to synthesize,[96] retrieving and manipulating information encoded in non-standard synthetic molecules may be limited by available techniques and instrumentation. In contrast, DNA-based storage takes advantage of rapidly evolving tools from life sciences, and we believe DNA is by far the most feasible material for building practical molecular storage systems in light of the technological readiness, achievable scale and robustness, as well as appealing properties such as programmable base pairing with rich potential for supporting molecular computations and database operations. Second, innovative strategies of mapping information to nucleotides could be devised to mitigate conventional synthesis challenges. Despite different encoding schemes, the majority of *in vitro* DNA storage demonstrated to date rely on *de novo* synthesis of large quantities of unique sequences. While coding density is a crucial consideration, such approach is unscalable and becomes costly as the data size increases. Alternatively, one could design and prefabricate a library of short DNA segments as codeword lexicon such that arbitrary data can be represented by rearranging the segments and concatenating them into addressable payload strands *via* efficient enzymatic reactions. With carefully optimized codeword designs to minimize spurious cross-talk and maximize the sequence diversity for data encoding flexibility, such modular approach may drastically improve the scalability and cost efficiency to pave the way for commercializing practical DNA storage.[97] Third, recent DNA storage with error-correcting schemes have demonstrated tolerance to synthesis and sequencing errors and even loss of complete sequences. This implies that affordable and low-fidelity technologies might be sufficient to support practical DNA storage if the encoding and error-correcting strategies could be further optimized. While we see plenty of room for improvement down this route, a considerable amount of work is needed including thorough evaluations of DNA storage error models, *etc.* Error tolerance could be readily enhanced by increasing the physical/logical redundancy of data stored in DNA, however, cost would be a significant tradeoff. Nevertheless, with joint efforts from academia and industry, we believe the opportunities are ample to tackle the challenges of building low-cost practical DNA storage in the foreseeable future.

As DNA could function as both a storage medium and a computing substrate, bridging DNA-based storage and computing may be the next important step toward constructing DNA computers to fully leverage the power of molecular parallelism in the context of near-data processing.[22] The field of DNA computing has progressed rapidly in recent years and various molecular computing frameworks have been devised based on elegant and versatile primitives such as DNA strand displacement. While DNA offers many benefits including biocompatibility, programmability, and massive molecular parallelism, DNA-based computing cannot match the speed and precision of typical electronics due to the chemical reaction kinetics. To address the challenge, strategies such as localized circuits[19,70] have been implemented with promising results. In prior DNA storage systems, operations such as PCR-based random access,[9,11–13] hybridization-based associative search,[78,98] and content-based similarity match[22,79] have all leveraged strand interactions. Following the trend, we envision that more sophisticated DNA storage features and operations could be engineered by strategically adapting and extending the knowledge from the field of computer science and

DNA nanotechnology. In fact, researchers have used programmable self-assembly of DNA nanostructures to construct multi-bit rewritable DNA memories *in vitro*.[99,100] By encoding bits in the form of strand interactions, these systems could easily leverage hybridization or strand displacement to support on-memory computing and parallel bit operations. Utilizing strand interactions for DNA database operations requires systematically designed storage and computing strands to reliably communicate and execute instructions. Therefore, the design and synthesis of the computing strands would face similar challenges as the storage strands, and additional data transduction modules may be necessary to properly interface storage with computing. Alternatively, it might be sufficient to execute computations exclusively at the primer level (*i.e.* on the addresses of storage strands) to simplify the architectural designs for static DNA databases with predetermined storage hierarchy. Moreover, implementing DNA database instructions such as insert, update, delete, and relational join is likely to involve both DNA and enzymes, which imposes constraints on sequence designs and necessitates systematic approaches for architecting modular and scalable systems. As it is desirable for database systems to support repeated data manipulations, the synthetic DNA molecules used in DNA databases should ideally be reusable for multiple cycles of operation. In addition to *in vitro* dynamic DNA circuits,[20] several schemes of renewable DNA reaction networks[64,65,101] have been proposed and may shed light on rational designs of reusable DNA storage/computing architectures. Researchers have also integrated DNA memory with electrode arrays[102] or fluorescence switching systems[103] on microchips to enable rapid random access, rewritability, and bitwise logical operations. These hybrid DNA storage systems could leverage the strengths of different technologies and offer more dimensions of control. For instance, in addition to PCR-based random access, magnetic beads[79] and microfluidic systems[104] can be utilized to accommodate new modalities of data retrieval from physically separated pools of DNA storage. Furthermore, we believe the investigation on unconventional paradigms such as amorphous computing[105] could offer critical insights to fully unleash the characteristic potential of molecular storage and computing in particularly large reaction networks. The requirements and challenges for bridging DNA computing with storage are different for *in vitro* versus *in vivo* applications. For *in vitro* systems such as DNA archival storage, major considerations include coding density, storage capacity and scalability, as well as tolerance to errors and degradation. For *in vivo* systems, criteria such as coding density may not be as critical because information could be collectively gathered from a large population of cells. *In vivo* DNA storage/computing applications instead focus more on the design of biosynthetic systems to properly sense from, interface with, and react to the biological environment to faithfully detect, interpret, and record events and signals of interest. Last but not least, it is important to assess the nontechnical implications involved in the development of DNA storage and computing systems. In particular, the biological relevance of DNA necessitates detailed regulations to guide the technology commercialization and maintain both digital and biological security. While challenges remain, the future of synthetic DNA storage/computing systems is bright and may profoundly impact areas such as global data management and health care.

Vocabulary

Directed spacer acquisition: the ordered integration of intracellularly generated or externally introduced synthetic protospacer sequences into genomic CRISPR arrays. **Error-correcting code:** techniques of adding data redundancy to allow error-free data reconstruction from unreliable data transmission or storage. Examples include the Hamming code, Reed-Solomon (RS) code, *etc.* **Nested PCR:** a modified PCR technique that leverages multiple sets of primers during sequential PCR cycles to enhance the specificity of the target DNA amplification. **Random access:** the ability to efficiently retrieve an arbitrary data element from a large population of addressable data items without sequentially traversing through all data locations. **Self-targeting CRISPR/Cas:** a modified CRISPR/Cas system that incorporates a PAM in the sgRNA-encoding locus such that the transcribed sgRNA repeatedly directs the Cas-mediated cleavage toward the sgRNA locus, leading to continuous mutagenesis and evolution of the target sequence.

Figures and Tables

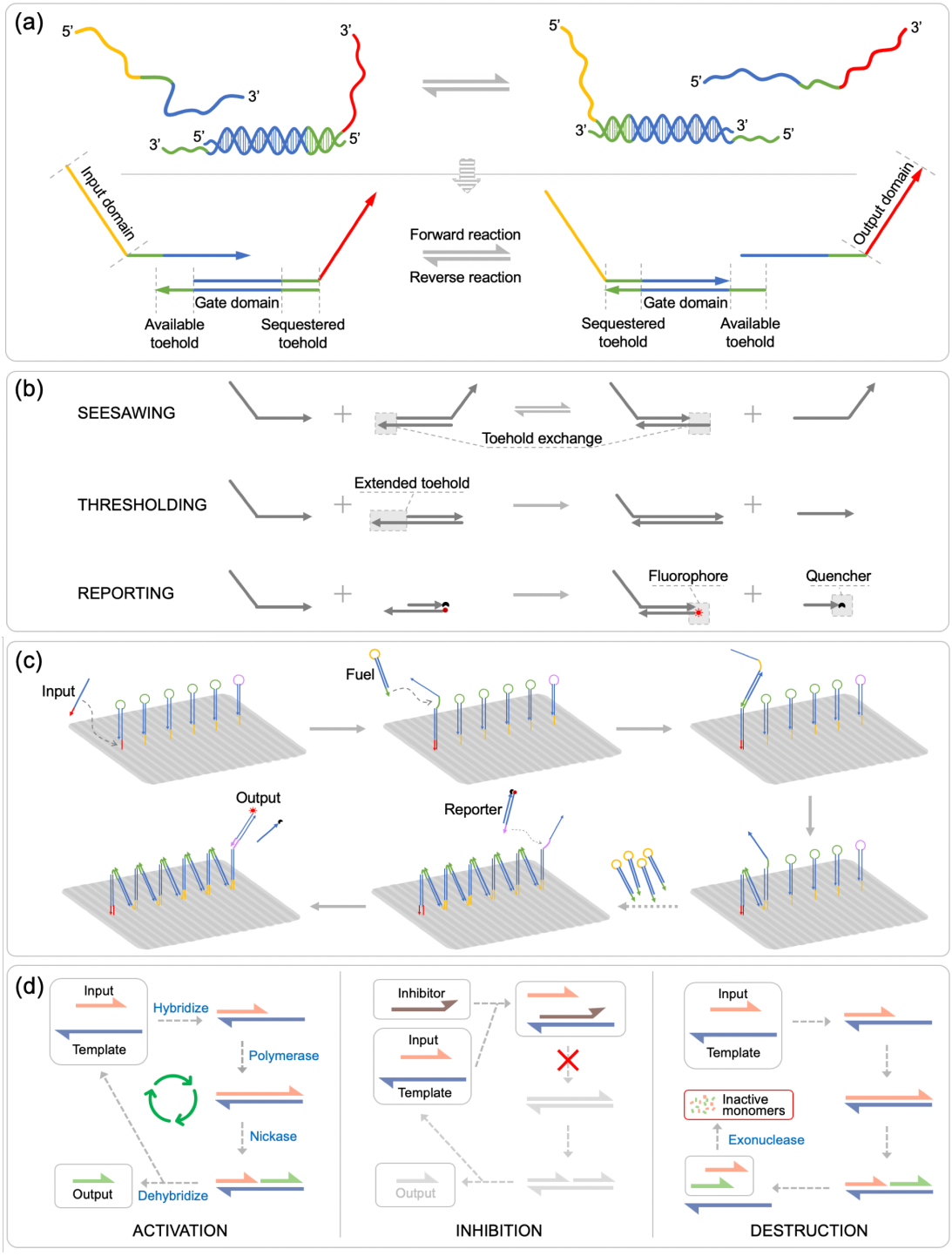


Figure 1. Example primitives for *in vitro* DNA computing. (a) Toehold-mediated strand displacement and domain-based sequence abstraction. (b) Basic reactions in DNA seesaw circuits. (c) Localized DNA hairpin hybridization chain reactions. (d) Basic modules for dynamic DNA reaction networks.

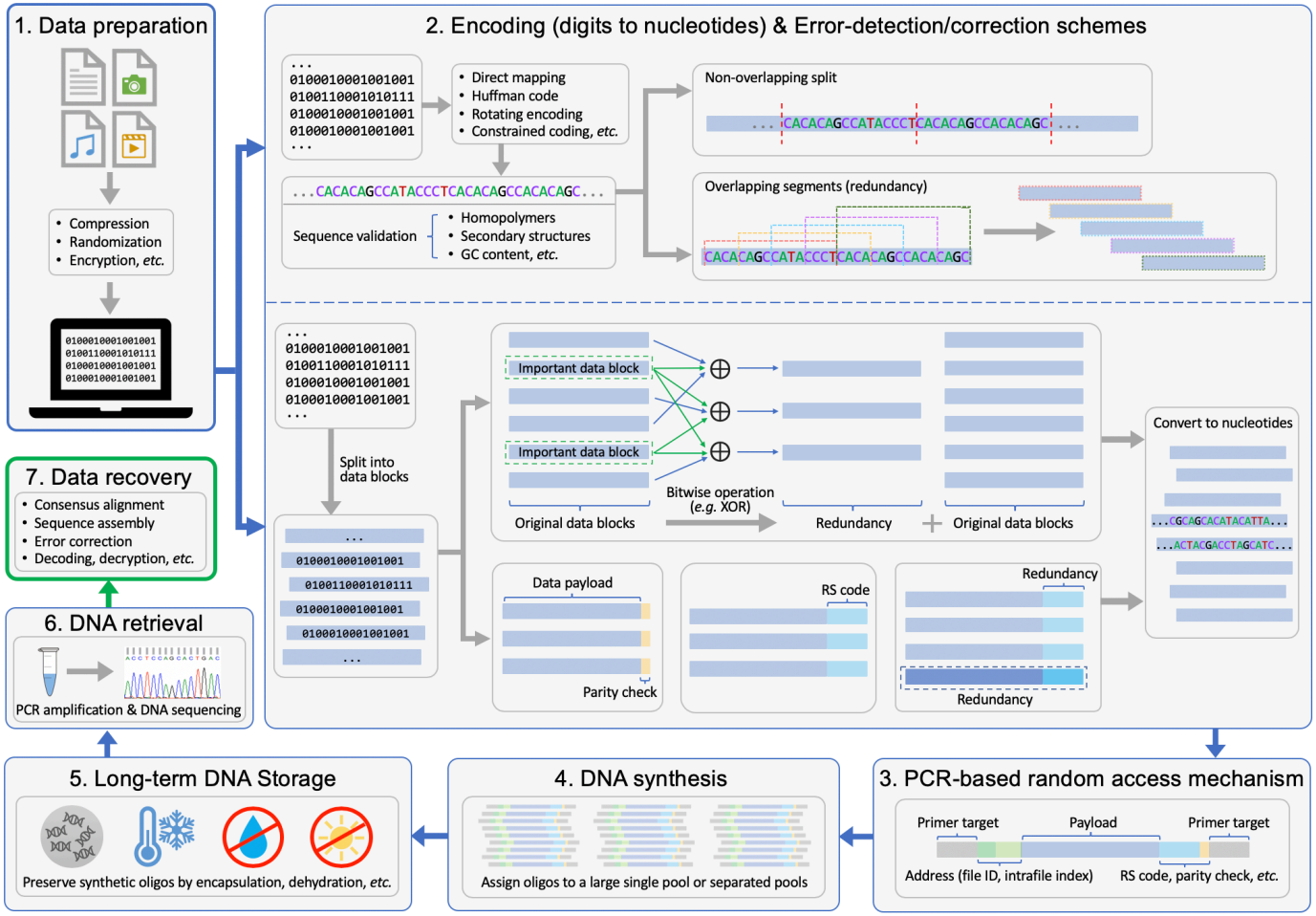


Figure 2. Example workflow for constructing large synthetic DNA databases. (1) Data files of various types are converted to binary digits after compression and encryption. (2) Digits are converted to nucleotides *via* different coding algorithms. Due to limits of synthesis/sequencing technology, data are typically split into small chunks. To enhance error tolerance, data redundancy and/or error correction codes can be introduced to the binary data chunks or to the encoded nucleotide segments. The resulting data-encoding oligos are screened in terms of GC contents and secondary structures *etc.* to reduce synthesis/sequencing errors. (3) Primers and addresses are added to data-encoding oligos to enable PCR-based random access. (4-5) The resulting oligos are synthesized and stored for long-term preservation. (6) To retrieve data stored in DNA, a sample of the DNA storage pool is PCR amplified using specific address primers. The resulting amplicons are then read out by sequencing. (7) After sequence alignment and assembly, the data can be decoded and reconstructed after error correction and decryption.

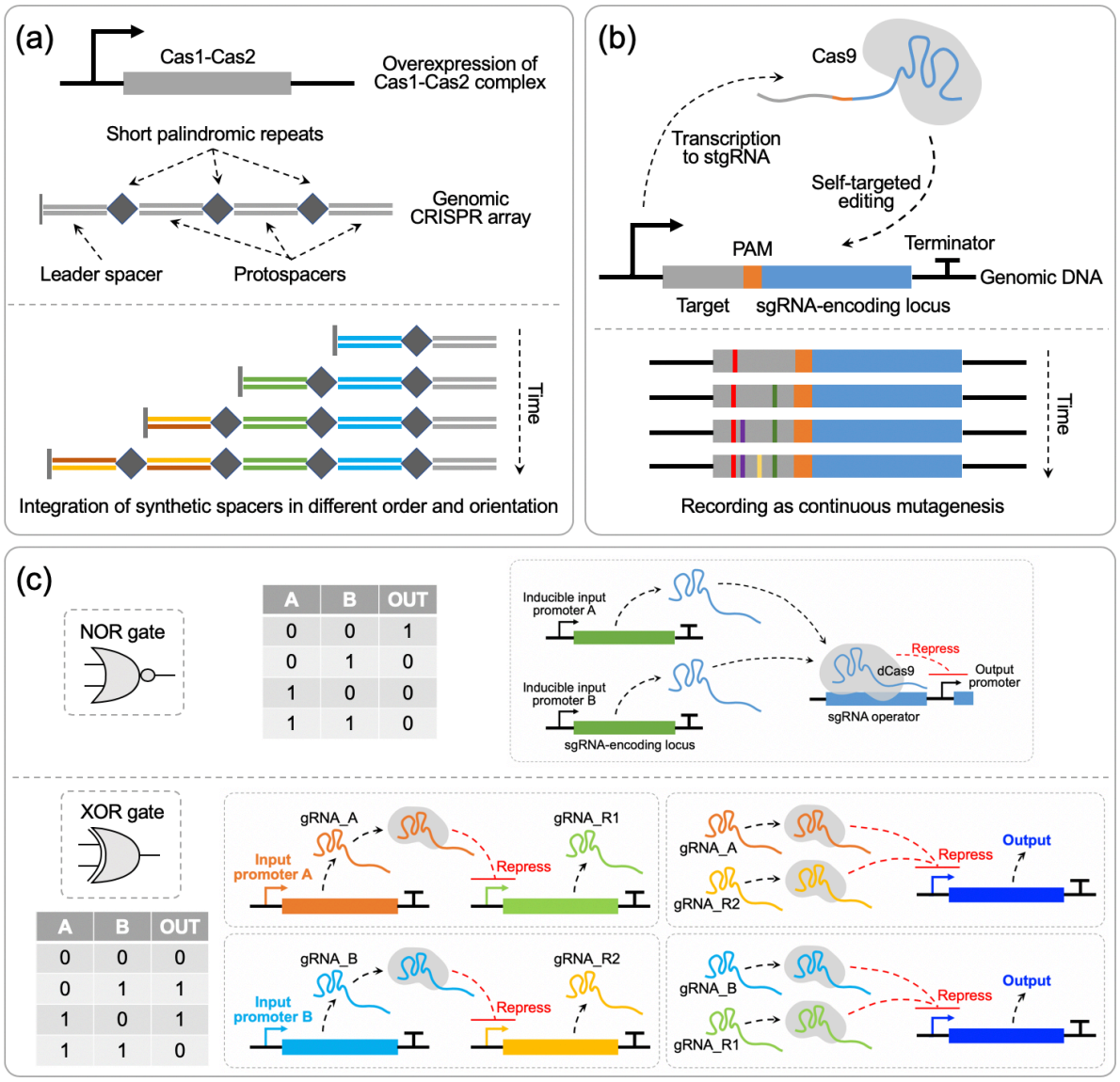


Figure 3. Examples of *in vivo* DNA memory and computing. (a) Temporal recording *via* directed protospacer acquisition in CRISPR arrays. (b) Continuous recording *via* self-targeting CRISPR/Cas. (c) Transcriptional logic gates by CRISPR/dCas9-mediated repression of synthetic promoters.

Table 1. Comparison of non-volatile storage media

Media type	Magnetic tape	Floppy disk	Hard disk drive	Flash memory [a]	Optical storage [b]	DNA storage
Advantages	Low cost, high capacity, transportable and scalable	Portable, random access	High capacity, random access, low cost	Fast random access, no moving parts, compact and lightweight	High capacity, low cost, easy to transport	Dense and durable, enormous capacity, low power consumption
Disadvantages	Slow linear access, susceptible to material degradation	Low capacity, low reliability, obsolete technology	Short lifespan, high power consumption, prone to errors/mechanical failure	Short lifespan, limited capacity, high cost	Moving mechanical parts, prone to damages	Slow read/write speed, specialized equipment, emerging technology
Applications	Backup and archival storage, audio/video recording	Data transfer between early computers	Computer data storage	Digital data transfer and storage	Audio/video file distribution, archival storage	Long-term data preservation, steganography

[a] Examples of flash memory include USB thumb drive, SD card, and solid-state drive. [b] Examples of optical storage media include CD, DVD, Blu-ray disc, and holographic disc.

Table 2. Comparison of recent implementations of large DNA databases

Publication	Church *et al.* (2012)[4]	Goldman *et al.* (2013)[5]	Yazdi *et al.* (2015)[11]	Grass *et al.* (2015)[6]	Bornholt *et al.* (2016)[12]	Blawat *et al.* (2016)[7]	Erlich and Zielinski (2017)[8]	Yazdi *et al.* (2017)[9]	Organick *et al.* (2018)[13]	
Stored data size	0.66 MB	739 KB	17 KB	83 KB	151 KB	22 MB	2.14 MB	3.63 KB	200.2 MB	33.3 KB
Net coding density (bits/nt)	0.83	0.33	N/A	1.14	0.85	1.03	1.57	1.72	1.10	
Error detection/correction	None	Redundancy (overlapping), parity check, majority voting	Prefix-synchronized encoding	Concatenated RS codes	Tunable redundancy (XOR), parity check	Alternating mapping, BCH code, RS code, parity check, majority voting	Fountain code, RS code	Consensus alignment, majority voting, homopolymer check	Input randomization, concatenated codes, similarity-based clustering, bitwise majority alignment	
Error-free data recovery	×	✓ (with manual intervention)	✓	✓	✓ (with manual intervention)	✓	✓	✓	✓	
Random access	×	×	✓	×	✓	×	×	✓	✓	
Content rewritability	×	×	✓	×	×	×	×	×	×	
Oligo pool size	54898	153335	32	4991	45652	900000	72000	17	13448372	2130
Oligo length	115 nt	117 nt	1000 bp	117 nt	120 nt	190 nt	152 nt	1000 bp	110 nt	N/A
Coverage for reliable reconstruction	3000x	519x, 52x (subsampled)	N/A	372x (untreated), 456x (1 week of thermal treatment)	122x (subsampled)	161x	444x, 10x (subsampled), 69x (deep copied)	392x	5x (subsampled)	36x, 80x (subsampled)
Sequencing platform	Illumina HiSeq	Illumina HiSeq	Sanger	Illumina MiSeq	Illumina MiSeq	Illumina HiSeq	Illumina MiSeq	Nanopore MinION	Illumina NextSeq	Nanopore MinION

Acknowledgments

This work was sponsored by NSF Grant No. CCF 1617791. X.S. acknowledges support from NSF Grant No. DGE 1545220.

References

(1) Goda, K.; Kitsuregawa, M. The History of Storage Systems. *Proc. IEEE* **2012**, *100*, 1433–1440.
(2) Greengard, S. Cracking the Code on DNA Storage. *Commun. ACM* **2017**, *60*, 16–18.
(3) Zhirnov, V.; Zadegan, R. M.; Sandhu, G. S.; Church, G. M.; Hughes, W. L. Nucleic Acid Memory. *Nat. Mater.* **2016**, *15*, 366–370.
(4) Church, G. M.; Gao, Y.; Kosuri, S. Next-Generation Digital Information Storage in DNA. *Science* **2012**, *337*, 1628–1628.
(5) Goldman, N.; Bertone, P.; Chen, S.; Dessimoz, C.; LeProust, E. M.; Sipos, B.; Birney, E. Towards Practical, High-Capacity, Low-Maintenance Information Storage in Synthesized DNA. *Nature* **2013**, *494*, 77–80.
(6) 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. Chemie Int. Ed.* **2015**, *54*, 2552–2555.
(7) Blawat, M.; Gaedke, K.; Hütter, I.; Chen, X.-M.; Turczyk, B.; Inverso, S.; Pruitt, B. W.; Church, G. M. Forward Error Correction for DNA Data Storage. *Procedia Comput. Sci.* **2016**, *80*, 1011–1022.
(8) Erlich, Y.; Zielinski, D. DNA Fountain Enables a Robust and Efficient Storage Architecture. *Science* **2017**, *355*, 950–954.
(9) Yazdi, S. M. H. T.; Gabrys, R.; Milenkovic, O. Portable and Error-Free DNA-Based Data Storage. *Sci. Rep.* **2017**, *7*, 5011.
(10) Heckel, R.; Shomorony, I.; Ramchandran, K.; Tse, D. N. C. Fundamental Limits of DNA Storage Systems. In *International Symposium on Information Theory (ISIT)*; IEEE: Aachen, 2017; pp 3130–3134.
(11) Yazdi, S. M. H. T.; Yuan, Y.; Ma, J.; Zhao, H.; Milenkovic, O. A Rewritable, Random-Access DNA-Based Storage System. *Sci. Rep.* **2015**, *5*, 14138.
(12) Bornholt, J.; Lopez, R.; Carmean, D. M.; Ceze, L.; Seelig, G.; Strauss, K. A DNA-Based Archival Storage System. *ACM SIGARCH Comput. Archit. News* **2016**, *44*, 637–649.
(13) Organick, L.; Ang, S. D.; Chen, Y.-J.; Lopez, R.; Yekhanin, S.; Makarychev, K.; Racz, M. Z.; Kamath, G.; Gopalan, P.; Nguyen, B.; Takahashi, C. N.; Newman, S.; Parker, H.-Y.; Rashtchian, C.; Stewart, K.; Gupta, G.; Carlson, R.; Mulligan, J.; Carmean, D.; Seelig, G.; *et al.* Random Access in Large-Scale DNA Data Storage. *Nat. Biotechnol.* **2018**, *36*, 242–248.
(14) Farzadfard, F.; Lu, T. K. Emerging Applications for DNA Writers and Molecular Recorders. *Science* **2018**, *361*, 870–875.
(15) Sheth, R. U.; Wang, H. H. DNA-Based Memory Devices for Recording Cellular Events. *Nat. Rev. Genet.* **2018**, *19*, 718–732.
(16) Adleman, L. Molecular Computation of Solutions to Combinatorial Problems. *Science* **1994**, *266*, 1021–1024.
(17) Qian, L.; Winfree, E. A Simple DNA Gate Motif for Synthesizing Large-Scale Circuits. *J. R. Soc. Interface* **2011**, *8*, 1281–1297.
(18) Cherry, K. M.; Qian, L. Scaling Up Molecular Pattern Recognition with DNA-Based Winner-Take-All Neural Networks. *Nature* **2018**, *559*, 370–376.
(19) Chatterjee, G.; Dalchau, N.; Muscat, R. A.; Phillips, A.; Seelig, G. A Spatially Localized Architecture for Fast and Modular DNA Computing. *Nat. Nanotechnol.* **2017**, *12*, 920–927.
(20) Montagne, K.; Plasson, R.; Sakai, Y.; Fujii, T.; Rondelez, Y. Programming an *In Vitro* DNA

Oscillator Using a Molecular Networking Strategy. *Mol. Syst. Biol.* **2011**, *7*, 466–466.
(21) Reif, J. H. Parallel Molecular Computation. In *Proceedings of the seventh annual ACM symposium on Parallel algorithms and architectures*; ACM Press: New York, New York, USA, 1995; pp 213–223.
(22) Carmean, D.; Ceze, L.; Seelig, G.; Stewart, K.; Strauss, K.; Willsey, M. DNA Data Storage and Hybrid Molecular–Electronic Computing. *Proc. IEEE* **2019**, *107*, 63–72.
(23) Qian, Y.; McBride, C.; Del Vecchio, D. Programming Cells to Work for Us. *Annu. Rev. Control. Robot. Auton. Syst.* **2018**, *1*, 411–440.
(24) Siuti, P.; Yazbek, J.; Lu, T. K. Synthetic Circuits Integrating Logic and Memory in Living Cells. *Nat. Biotechnol.* **2013**, *31*, 448–452.
(25) Yehl, K.; Lu, T. Scaling Computation and Memory in Living Cells. *Curr. Opin. Biomed. Eng.* **2017**, *4*, 143–151.
(26) Pääbo, S.; Gifford, J. A.; Wilson, A. C. Mitochondrial DNA Sequences from a 7000-Year Old Brain. *Nucleic Acids Res.* **1988**, *16*, 9775–9787.
(27) Vreeland, R. H.; Rosenzweig, W. D.; Powers, D. W. Isolation of a 250 Million-Year-Old Halotolerant Bacterium from a Primary Salt Crystal. *Nature* **2000**, *407*, 897–900.
(28) Anchordoquy, T. J.; Molina, M. C. Preservation of DNA. *Cell Preserv. Technol.* **2007**, *5*, 180–188.
(29) Paunescu, D.; Fuhrer, R.; Grass, R. N. Protection and Deprotection of DNA-High-Temperature Stability of Nucleic Acid Barcodes for Polymer Labeling. *Angew. Chemie Int. Ed.* **2013**, *52*, 4269–4272.
(30) Zuker, M. Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction. *Nucleic Acids Res.* **2003**, *31*, 3406–3415.
(31) Zadeh, J. N.; Steenberg, C. D.; Bois, J. S.; Wolfe, B. R.; Pierce, M. B.; Khan, A. R.; Dirks, R. M.; Pierce, N. A. NUPACK: Analysis and Design of Nucleic Acid Systems. *J. Comput. Chem.* **2011**, *32*, 170–173.
(32) Snodin, B. E. K.; Randisi, F.; Mosayebi, M.; Šulc, P.; Schreck, J. S.; Romano, F.; Ouldridge, T. E.; Tsukanov, R.; Nir, E.; Louis, A. A.; Doye, J. P. K. Introducing Improved Structural Properties and Salt Dependence into a Coarse-Grained Model of DNA. *J. Chem. Phys.* **2015**, *142*, 234901.
(33) Zhang, D. Y. Towards Domain-Based Sequence Design for DNA Strand Displacement Reactions. In *DNA Computing and Molecular Programming. DNA 2010. Lecture Notes in Computer Science*; Sakakibara, Y., Mi, Y., Eds.; Springer: Berlin, Heidelberg, 2011; Vol. 6518, pp 162–175.
(34) Hughes, R. A.; Ellington, A. D. Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. *Cold Spring Harb. Perspect. Biol.* **2017**, *9*, a023812.
(35) Service, R. F. DNA Printers Poised to Jump from Paragraphs to Pages. *Science.* **2018**, *362*, 143–143.
(36) Mardis, E. R. DNA Sequencing Technologies: 2006–2016. *Nat. Protoc.* **2017**, *12*, 213–218.
(37) Garibyan, L.; Avashia, N. Research Techniques Made Simple: Polymerase Chain Reaction (PCR). *J. Invest. Dermatol.* **2013**, *133*, 1–4.
(38) Yamamoto, M.; Kashiwamura, S.; Ohuchi, A.; Furukawa, M. Large-Scale DNA Memory Based on the Nested PCR. *Nat. Comput.* **2008**, *7*, 335–346.
(39) Harrison, P. T.; Hart, S. A Beginner's Guide to Gene Editing. *Exp. Physiol.* **2018**, *103*, 439–448.
(40) Gaj, T.; Gersbach, C. A.; Barbas, C. F. ZFN, TALEN, and CRISPR/Cas-Based Methods for Genome Engineering. *Trends Biotechnol.* **2013**, *31*, 397–405.
(41) Barrangou, R.; Fremaux, C.; Deveau, H.; Richards, M.; Boyaval, P.; Moineau, S.; Romero, D. A.; Horvath, P. CRISPR Provides Acquired Resistance against Viruses in Prokaryotes. *Science* **2007**, *315*, 1709–1712.
(42) Cong, L.; Ran, F. A.; Cox, D.; Lin, S.; Barretto, R.; Habib, N.; Hsu, P. D.; Wu, X.; Jiang, W.; Marraffini, L. A.; Zhang, F. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **2013**, *339*, 819–823.
(43) Shalem, O.; Sanjana, N. E.; Hartenian, E.; Shi, X.; Scott, D. A.; Mikkelsen, T. S.; Heckl, D.;

Ebert, B. L.; Root, D. E.; Doench, J. G.; Zhang, F. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science* **2014**, *343*, 84–87.
(44) Komor, A. C.; Kim, Y. B.; Packer, M. S.; Zuris, J. A.; Liu, D. R. Programmable Editing of a Target Base in Genomic DNA Without Double-Stranded DNA Cleavage. *Nature* **2016**, *533*, 420–424.
(45) Gaudelli, N. M.; Komor, A. C.; Rees, H. A.; Packer, M. S.; Badran, A. H.; Bryson, D. I.; Liu, D. R. Programmable Base Editing of A•T to G•C in Genomic DNA Without DNA Cleavage. *Nature* **2017**, *551*, 464–471.
(46) Barrangou, R.; Doudna, J. A. Applications of CRISPR Technologies in Research and Beyond. *Nat. Biotechnol.* **2016**, *34*, 933–941.
(47) Adli, M. The CRISPR Tool Kit for Genome Editing and Beyond. *Nat. Commun.* **2018**, *9*, 1911.
(48) Shipman, S. L.; Nivala, J.; Macklis, J. D.; Church, G. M. Molecular Recordings by Directed CRISPR Spacer Acquisition. *Science* **2016**, *353*, aaf1175.
(49) Shipman, S. L.; Nivala, J.; Macklis, J. D.; Church, G. M. CRISPR–Cas Encoding of a Digital Movie into the Genomes of a Population of Living Bacteria. *Nature* **2017**, *547*, 345–349.
(50) Sheth, R. U.; Yim, S. S.; Wu, F. L.; Wang, H. H. Multiplex Recording of Cellular Events over Time on CRISPR Biological Tape. *Science* **2017**, *358*, 1457–1461.
(51) Perli, S. D.; Cui, C. H.; Lu, T. K. Continuous Genetic Recording with Self-Targeting CRISPR-Cas in Human Cells. *Science* **2016**, *353*, aag0511–aag0511.
(52) Kalhor, R.; Mali, P.; Church, G. M. Rapidly Evolving Homing CRISPR Barcodes. *Nat. Methods* **2017**, *14*, 195–200.
(53) Jusiak, B.; Cleto, S.; Perez-Piñera, P.; Lu, T. K. Engineering Synthetic Gene Circuits in Living Cells with CRISPR Technology. *Trends Biotechnol.* **2016**, *34*, 535–547.
(54) Zhang, D. Y.; Seelig, G. Dynamic DNA Nanotechnology Using Strand-Displacement Reactions. *Nat. Chem.* **2011**, *3*, 103–113.
(55) Lakin, M. R.; Youssef, S.; Polo, F.; Emmott, S.; Phillips, A. Visual DSD: A Design and Analysis Tool for DNA Strand Displacement Systems. *Bioinformatics* **2011**, *27*, 3211–3213.
(56) Salehi, S. A.; Liu, X.; Riedel, M. D.; Parhi, K. K. Computing Mathematical Functions Using DNA *via* Fractional Coding. *Sci. Rep.* **2018**, *8*, 8312.
(57) Seelig, G.; Soloveichik, D.; Zhang, D. Y.; Winfree, E. Enzyme-Free Nucleic Acid Logic Circuits. *Science* **2006**, *314*, 1585–1588.
(58) Chen, Y. J.; Dalchau, N.; Srinivas, N.; Phillips, A.; Cardelli, L.; Soloveichik, D.; Seelig, G. Programmable Chemical Controllers Made from DNA. *Nat. Nanotechnol.* **2013**, *8*, 755–762.
(59) Yurke, B.; Turberfield, A. J.; Mills, A. P.; Simmel, F. C.; Neumann, J. L. A DNA-Fuelled Molecular Machine Made of DNA. *Nature* **2000**, *406*, 605–608.
(60) Zhang, D. Y.; Turberfield, A. J.; Yurke, B.; Winfree, E. Engineering Entropy-Driven Reactions and Networks Catalyzed by DNA. *Science* **2007**, *318*, 1121–1125.
(61) Zhang, D. Y.; Winfree, E. Control of DNA Strand Displacement Kinetics Using Toehold Exchange. *J. Am. Chem. Soc.* **2009**, *131*, 17303–17314.
(62) Qian, L.; Winfree, E. Scaling up Digital Circuit Computation with DNA Strand Displacement Cascades. *Science* **2011**, *332*, 1196–1201.
(63) Qian, L.; Winfree, E.; Bruck, J. Neural Network Computation with DNA Strand Displacement Cascades. *Nature* **2011**, *475*, 368–372.
(64) Song, X.; Eshra, A.; Dwyer, C.; Reif, J. Renewable DNA Seesaw Logic Circuits Enabled by Photoregulation of Toehold-Mediated Strand Displacement. *RSC Adv.* **2017**, *7*, 28130–28144.
(65) Garg, S.; Shah, S.; Bui, H.; Song, T.; Mokhtar, R.; Reif, J. Renewable Time-Responsive DNA Circuits. *Small* **2018**, *14*, 1801470.
(66) Wilhelm, D.; Bruck, J.; Qian, L. Probabilistic Switching Circuits in DNA. *Proc. Natl. Acad. Sci.* **2018**, *115*, 903–908.
(67) Chandran, H.; Gopalkrishnan, N.; Phillips, A.; Reif, J. Localized Hybridization Circuits. In *DNA Computing and Molecular Programming. DNA 2011. Lecture Notes in Computer Science*;

Cardelli, L., Shih, W., Eds.; Springer: Berlin, Heidelberg, 2011; Vol. 6937, pp 64–83.
(68) Dalchau, N.; Chandran, H.; Gopalkrishnan, N.; Phillips, A.; Reif, J. Probabilistic Analysis of Localized DNA Hybridization Circuits. *ACS Synth. Biol.* **2015**, *4*, 898–913.
(69) Bui, H.; Miao, V.; Garg, S.; Mokhtar, R.; Song, T.; Reif, J. Design and Analysis of Localized DNA Hybridization Chain Reactions. *Small* **2017**, *13*, 1602983.
(70) Bui, H.; Shah, S.; Mokhtar, R.; Song, T.; Garg, S.; Reif, J. Localized DNA Hybridization Chain Reactions on DNA Origami. *ACS Nano* **2018**, *12*, 1146–1155.
(71) Engelen, W.; Wijnands, S. P. W.; Merkx, M. Accelerating DNA-Based Computing on a Supramolecular Polymer. *J. Am. Chem. Soc.* **2018**, *140*, 9758–9767.
(72) Youn Lee, J.; Shin, S. Y.; June Augh, S.; Hyun Park, T.; Zhang, B. T. Temperature Gradient-Based DNA Computing for Graph Problems with Weighted Edges. In *DNA 2002. Lecture Notes in Computer Science*; Hagiya, M., Ohuchi, A., Eds.; Springer: Berlin, Heidelberg, 2003; Vol. 2568, pp 73–84.
(73) Sakamoto, K. Molecular Computation by DNA Hairpin Formation. *Science* **2000**, *288*, 1223–1226.
(74) Komiya, K.; Sakamoto, K.; Kameda, A.; Yamamoto, M.; Ohuchi, A.; Kiga, D.; Yokoyama, S.; Hagiya, M. DNA Polymerase Programmed with a Hairpin DNA Incorporates a Multiple-Instruction Architecture into Molecular Computing. *Biosystems* **2006**, *83*, 18–25.
(75) Yang, J.; Wu, R.; Li, Y.; Wang, Z.; Pan, L.; Zhang, Q.; Lu, Z.; Zhang, C. Entropy-Driven DNA Logic Circuits Regulated by DNAzyme. *Nucleic Acids Res.* **2018**, *46*, 8532–8541.
(76) Elbaz, J.; Lioubashevski, O.; Wang, F.; Remacle, F.; Levine, R. D.; Willner, I. DNA Computing Circuits Using Libraries of DNAzyme Subunits. *Nat. Nanotechnol.* **2010**, *5*, 417–422.
(77) Baum, E. Building an Associative Memory Vastly Larger than the Brain. *Science* **1995**, *268*, 583–585.
(78) Reif, J. H.; LaBean, T. H.; Pirrung, M.; Rana, V. S.; Guo, B.; Kingsford, C.; Wickham, G. S. Experimental Construction of Very Large Scale DNA Databases with Associative Search Capability. In *DNA Computing. DNA 2001. Lecture Notes in Computer Science*; Jonoska, N., Seeman, N. C., Eds.; Springer: Berlin, Heidelberg, 2002; Vol. 2340, pp 231–247.
(79) Stewart, K.; Chen, Y.-J.; Ward, D.; Liu, X.; Seelig, G.; Strauss, K.; Ceze, L. A Content-Addressable DNA Database with Learned Sequence Encodings. In *DNA Computing and Molecular Programming. DNA 2018. Lecture Notes in Computer Science*; Doty, D., Dietz, H., Eds.; Springer: Cham, 2018; Vol. 11145, pp 55–70.
(80) Ham, T. S.; Lee, S. K.; Keasling, J. D.; Arkin, A. P. Design and Construction of a Double Inversion Recombination Switch for Heritable Sequential Genetic Memory. *PLoS One* **2008**, *3*, e2815.
(81) Bonnet, J.; Subsoontorn, P.; Endy, D. Rewritable Digital Data Storage in Live Cells *via* Engineered Control of Recombination Directionality. *Proc. Natl. Acad. Sci.* **2012**, *109*, 8884–8889.
(82) Yang, L.; Nielsen, A. A. K.; Fernandez-Rodriguez, J.; McClune, C. J.; Laub, M. T.; Lu, T. K.; Voigt, C. A. Permanent Genetic Memory with >1-Byte Capacity. *Nat. Methods* **2014**, *11*, 1261–1266.
(83) Roquet, N.; Soleimany, A. P.; Ferris, A. C.; Aaronson, S.; Lu, T. K. Synthetic Recombinase-Based State Machines in Living Cells. *Science* **2016**, *353*, aad8559–aad8559.
(84) Farzadfard, F.; Lu, T. K. Genomically Encoded Analog Memory with Precise *In Vivo* DNA Writing in Living Cell Populations. *Science* **2014**, *346*, 1256272–1256272.
(85) Frieda, K. L.; Linton, J. M.; Hormoz, S.; Choi, J.; Chow, K.-H. K.; Singer, Z. S.; Budde, M. W.; Elowitz, M. B.; Cai, L. Synthetic Recording and *In Situ* Readout of Lineage Information in Single Cells. *Nature* **2017**, *541*, 107–111.
(86) Tang, W.; Liu, D. R. Rewritable Multi-Event Analog Recording in Bacterial and Mammalian Cells. *Science* **2018**, *360*, eaap8992.
(87) Nielsen, A. A.; Voigt, C. A. Multi-Input CRISPR/Cas Genetic Circuits That Interface Host

Regulatory Networks. *Mol. Syst. Biol.* **2014**, *10*, 763–763.
(88) Kim, H.; Bojar, D.; Fussenegger, M. A CRISPR/Cas9-Based Central Processing Unit to Program Complex Logic Computation in Human Cells. *Proc. Natl. Acad. Sci.* **2019**, *116*, 7214–7219.
(89) Farzadfard, F.; Gharaei, N.; Higashikuni, Y.; Jung, G.; Cao, J.; Lu, T. K. Single-Nucleotide-Resolution Computing and Memory in Living Cells. *bioRxiv* **2018**.
(90) Clelland, C. T.; Risca, V.; Bancroft, C. Hiding Messages in DNA Microdots. *Nature* **1999**, *399*, 533–534.
(91) Leier, A.; Richter, C.; Banzhaf, W.; Rauhe, H. Cryptography with DNA Binary Strands. *Biosystems* **2000**, *57*, 13–22.
(92) Halvorsen, K.; Wong, W. P. Binary DNA Nanostructures for Data Encryption. *PLoS One* **2012**, *7*, e44212.
(93) Gehani, A.; LaBean, T.; Reif, J. DNA-Based Cryptography. In *Aspects of Molecular Computing. Lecture Notes in Computer Science*; Jonoska, N., Păun, G., Rozenberg, G., Eds.; Springer: Berlin, Heidelberg, 2003; Vol. 2950, pp 167–188.
(94) Heider, D.; Barnekow, A. DNA-Based Watermarks Using the DNA-Crypt Algorithm. *BMC Bioinformatics* **2007**, *8*, 176.
(95) Palluk, S.; Arlow, D. H.; de Rond, T.; Barthel, S.; Kang, J. S.; Bector, R.; Baghdassarian, H. M.; Truong, A. N.; Kim, P. W.; Singh, A. K.; Hillson, N. J.; Keasling, J. D. *De Novo* DNA Synthesis Using Polymerase-Nucleotide Conjugates. *Nat. Biotechnol.* **2018**, *36*, 645–650.
(96) Al Ouahabi, A.; Amalian, J.-A.; Charles, L.; Lutz, J.-F. Mass Spectrometry Sequencing of Long Digital Polymers Facilitated by Programmed Inter-Byte Fragmentation. *Nat. Commun.* **2017**, *8*, 967.
(97) Molteni, M. The rise of DNA data storage https://www.wired.com/story/the-rise-of-dna-data-storage/ (accessed Nov 17, 2018).
(98) Reif, J. H.; LaBean, T. H. Computationally Inspired Biotechnologies: Improved DNA Synthesis and Associative Search Using Error-Correcting Codes and Vector-Quantization? In *DNA Computing. DNA 2000. Lecture Notes in Computer Science*; Condon, A., Rozenberg, G., Eds.; Springer: Berlin, Heidelberg, 2001; Vol. 2054, pp 145–172.
(99) Takinoue, M.; Suyama, A. Hairpin-DNA Memory Using Molecular Addressing. *Small* **2006**, *2*, 1244–1247.
(100) Chandrasekaran, A. R.; Levchenko, O.; Patel, D. S.; MacIsaac, M.; Halvorsen, K. Addressable Configurations of DNA Nanostructures for Rewritable Memory. *Nucleic Acids Res.* **2017**, *45*, 11459–11465.
(101) Goel, A.; Ibrahimi, M. A Renewable, Modular, and Time-Responsive DNA Circuit. *Nat. Comput.* **2011**, *10*, 467–485.
(102) Song, Y.; Kim, S.; Heller, M. J.; Huang, X. DNA Multi-Bit Non-Volatile Memory and Bit-Shifting Operations Using Addressable Electrode Arrays and Electric Field-Induced Hybridization. *Nat. Commun.* **2018**, *9*, 281.
(103) Nguyen, H. H.; Park, J.; Hwang, S.; Kwon, O. S.; Lee, C.-S.; Shin, Y.-B.; Ha, T. H.; Kim, M. On-Chip Fluorescence Switching System for Constructing a Rewritable Random Access Data Storage Device. *Sci. Rep.* **2018**, *8*, 337.
(104) Newman, S.; Stephenson, A. P.; Willsey, M.; Nguyen, B. H.; Takahashi, C. N.; Strauss, K.; Ceze, L. High Density DNA Data Storage Library *via* Dehydration with Digital Microfluidic Retrieval. *Nat. Commun.* **2019**, *10*, 1706.
(105) Chen, X.; Ellington, A. D. Shaping up Nucleic Acid Computation. *Curr. Opin. Biotechnol.* **2010**, *21*, 392–400.