Biological life is one of the most complex and dynamic systems in nature. Through evolution and natural selection, vast biochemical and biological diversity has emerged, from complex molecules to multicellular life. These multiscale biological systems precisely generate and respond to a myriad of biotic signals of varying order and magnitude. Signals can take the form of ions, metabolites, nucleic acids or proteins, producing biochemical gradients and signalling cascades that propagate across many length and time scales within cells and across populations. The integration of these signals through genetic and epigenetic regulation at the transcriptional, translational and post-translational levels results in robust cellular behaviours. The spatiotemporal delineation and chronology of these biological signals and cellular states is thus paramount to our understanding of the fundamental organizing principles of biology.

Tracking multiple biological events simultaneously over time remains a challenge given the sheer number and diversity of signals present within a cell at any given moment. Quantifying these signals and processes in their native cellular and environmental context, which is often inaccessible, poses further practical and technical difficulties. Cellular information can currently be measured by a plethora of methodologies, each with their strengths and weaknesses. In the emerging genomic era, where DNA can be readily analysed and altered, new modalities of DNA-based cellular recording are poised to overcome these traditional limitations in biological information storage and analysis in a variety of settings.

DNA is the fundamental molecule by which information is stored and utilized to produce life. DNA is a high-density storage medium that can be quickly copied by exponential PCR amplification and stably preserved for decades to millennia. Biological information encoded in DNA can be directly converted into actionable cellular responses through gene regulation and expression. Although DNA is often thought of as a long-term information-bearing molecule, there are many examples of biological information storage and access through DNA within a single life cycle of an organism. Examples include phase variation, CRISPR-mediated immunity, mammalian adaptive immune systems, diversity-generating retroelements and programmed genome rearrangements. Advances in next-generation sequencing (NGS) and nucleic acid synthesis have ushered in a new era of rapid and inexpensive DNA reading and writing, which has further elevated the relevance of DNA as a meaningful information storage medium.

In this Review, we discuss recent progress in the emerging field of DNA-based recording technologies in living cells. We highlight key elements of biological information storage, suggest quantitative metrics to assess different recording approaches and outline technical challenges and knowledge gaps that still need to be addressed. We end by offering possible applications of DNA-based cellular recording and speculate on the future of this exciting area of research and development. Although epigenetic mechanisms, both molecular and cellular, such as protein-based feedback circuits, DNA methylation, chromatin conformation, prion states and neuronal networks, are clearly interesting and important modes of biological information transmission and storage, they are beyond the scope of this focused Review. For technologies that employ DNA barcodes for lineage tracing applications, we direct the readers to a recent in-depth review.

**DNA-based memory devices for recording cellular events**

Ravi U. Sheth and Harris H. Wang

Abstract | Measuring biological data across time and space is critical for understanding complex biological processes and for various biosurveillance applications. However, such data are often inaccessible or difficult to directly obtain. Less invasive, more robust and higher-throughput biological recording tools are needed to profile cells and their environments. DNA-based cellular recording is an emerging and powerful framework for tracking intracellular and extracellular biological events over time across living cells and populations. Here, we review and assess DNA recorders that utilize CRISPR nucleases, integrases and base-editing strategies, as well as recombinase and polymerase-based methods. Quantitative characterization, modelling and evaluation of these DNA-recording modalities can guide their design and implementation for specific application areas.

1Department of Systems Biology, Columbia University Medical Center, New York, NY, USA.
2Integrated Program in Cellular, Molecular and Biomedical Studies, Columbia University, New York, NY, USA.
3Department of Pathology and Cell Biology, Columbia University Medical Center, New York, NY, USA.
*e-mail: hw2429@columbia.edu

https://doi.org/10.1038/s41576-018-0052-8
Different types of biological data can be acquired from cells, including identity, quantity, spatial and temporal information. DNA, RNA, proteins and organic/inorganic metabolites constitute the key classes of molecules that are often measured when analyzing biological processes. In addition to these molecular data, cellular phenotypes that are reflective of more global states, such as growth rate, membrane permeability, electrophysiological gradients and oxidative stress, can also be quantified. Contemporary approaches to measure these biological data suffer from three key limitations (see the figure).

Many environments are difficult to access for direct measurement (for example, the gut or brain). Methods that require destructive processing steps (for example, cell or tissue fixation) cannot yield temporal biological data. Methods that allow continuous data acquisition, such as live-cell imaging, require direct access to the biological sample and instruments that cannot be miniaturized down to cellular sizes. Furthermore, the multiplexing capacity of most current methods is either limited or requires considerable disruption to the biological state of cells. In theory, DNA-based recording systems could overcome many of these challenges when deployed in live cells to store biological data into a permanent DNA medium over time for analysis at a later point.

### Contemporary limits of biological recording

**Box 1**

Different types of biological data can be acquired from cells, including identity, quantity, spatial and temporal information. DNA, RNA, proteins and organic/inorganic metabolites constitute the key classes of molecules that are often measured when analyzing biological processes. In addition to these molecular data, cellular phenotypes that are reflective of more global states, such as growth rate, membrane permeability, electrophysiological gradients and oxidative stress, can also be quantified. Contemporary approaches to measure these biological data suffer from three key limitations (see the figure).

Many environments are difficult to access for direct measurement (for example, the gut or brain). Methods that require destructive processing steps (for example, cell or tissue fixation) cannot yield temporal biological data. Methods that allow continuous data acquisition, such as live-cell imaging, require direct access to the biological sample and instruments that cannot be miniaturized down to cellular sizes. Furthermore, the multiplexing capacity of most current methods is either limited or requires considerable disruption to the biological state of cells. In theory, DNA-based recording systems could overcome many of these challenges when deployed in live cells to store biological data into a permanent DNA medium over time for analysis at a later point.

### Contemporary challenges in biological recording

<table>
<thead>
<tr>
<th>Not accessible to direct measurement</th>
<th>Cannot record over time</th>
<th>Require continuous data acquisition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Live cell imaging (e.g. GFP tags)</td>
<td><strong>Signal</strong></td>
<td><strong>Time</strong></td>
</tr>
<tr>
<td>Cellular biosensors (e.g. genetically encoded Ca²⁺ indicator)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imaging fixed samples (e.g. FISH)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immunoassays (e.g. ELISA)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcriptomics (e.g. RNA-seq)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mass spectrometry (MALDI-TOF)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA-based recording systems</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ELISA, enzyme-linked immunosorbent assay; FISH, fluorescence in situ hybridization; MALDI-TOF, matrix-assisted laser desorption/ionization-time of flight; RNA-seq, RNA sequencing.

### Strategies for DNA-based memory in cells

A universal information recording and storage system requires several essential elements: first, transformation of the information of interest into a standardized data format or data stream; second, recording the data into a physical medium; and third, conversion of the stored data back to a desired form that can be interpreted by the user or utilized by another system. A biological DNA-based version of such a memory system must also possess these key capacities (FIG. 1). First, information within a cell such as the presence of a metabolite or expression of a gene must be transformed into a format that is compatible with the recording system (for example, a biological signal that induces the expression of recording components). Next, this information must be written directly into DNA by alteration, deletion or addition of bases through various DNA-modifying enzymes, such as nucleases, integrases or recombinases. Finally, the stored data are read back out from the DNA using a multitude of techniques such as sequencing or imaging. The stored information can be further used to directly actuate or elicit a specific set of biological responses, such as gene expression. Below, we delineate each of these components and their implementation in contemporary DNA-based data recording and storage systems (TABLE 1).

### Signal detection and transformation

**Signal and input types.** Although there are a variety of biological signals present within a cell, the dynamic regulation of gene expression through transcription of mRNA is one of the most important and prevalently measured classes of cellular signals. The ensemble of transcription levels across all of its genes can represent a simplified ‘state’ of a cell. Beyond transcriptional states, proteins and metabolites, both intracellular and extracellular, represent other classes of cellular signals that can change during cellular growth, development and maintenance in different environments. Both the concentration and identities of these molecules can serve as inputs into a biological recording system. Finally, physiologically important characteristics of the intracellular and extracellular environment such as temperature, pH, oxidative stress, radiation levels or electrochemical and electromagnetic gradients can also be inputs for sensing and recording. For all of these input types, the presence or absence of the signal (digital state) and its intensity or magnitude (analog state) are important recordable information, as is their variation across space and time.

**Signal sensing.** Cells possess numerous native mechanisms to assess transcriptional states that can be co-opted for cellular memory devices. For instance, the transcription level of a gene of interest can be measured by linking its upstream promoter to a recording system to capture transient regulatory changes. Indeed, early bacterial gene expression screens utilized a strategy in which native promoters are fused to a recombinase-based reporter that permanently altered a genomic site to identify virulence pathways. Recording certain combinations of genes and their expression levels can capture even more complex cellular phenotypes of interest such as growth rate or cellular burden.

The levels of intracellular and extracellular chemical, metabolite, RNA or protein-based signals can be detected with a growing toolbox of engineered biosensors with high signal specificity. These modular sensors can convert a myriad of signal types such as cancer-associated antigens, pathogen-derived peptides,
Many sensors, such as transcription factors, two-component systems and more complex signalling cascades, couple binding of an input ligand to a sensory protein with altered transcription from a specific output promoter, which can then be readily linked to recording systems. Alternatively, RNA-based sensors such as RNA aptamers and riboswitches recognize specific metabolites and alter...
expression of an output gene by diverse mechanisms (for example, tuning of translation)\(^3\). Beyond chemical and protein ligands, RNA signals such as mRNA levels of endogenous genes or microRNAs can also be sensed via riboregulators, which bind target RNA molecules and alter expression of an output\(^29,30\).

**Signal transformation.** Once sensed, a signal of interest must be converted into a format that is capable of specifically activating a recording system. For many systems, this step simply involves expression of the recording machinery to mediate DNA modification. Alternatively, a transformation of the input signal into a different format may be required. For example, a transcriptional signal can be converted to an altered abundance of intracellular DNA by using a copy-number-inducible plasmid system, which subsequently is recorded into genomic arrays by Cas1-Cas2 CRISPR integrate systems as short spacers\(^1\). Signal transformation can be represented as a transfer function of signal input to the resulting recording activity; its detection threshold, dynamic range and response characteristics (analog versus digital) must match the desired application. Synthetic biology and genetic engineering techniques can be utilized to rationally alter and optimize this transformation (for instance, by tuning expression levels of recording machinery or altering sensor detection thresholds by protein engineering)\(^32\).

Synthetic gene circuits can be interfaced with biosensors for more complex tuning of signal transformation or to add more sophisticated functionality such as signal integration and computation\(^27,32\). For example, signal processing circuits can be linked to biosensors to achieve digital or analog responses to an input signal\(^14,38\). In order to alter signal response dynamics and record rapidly fluctuating signals, positive feedback and memory modules can be utilized\(^17\). In more complex eukaryotic signalling cascades, scaffold proteins can be shuffled or modified to achieve digital or analog responses to a signal\(^14,38\).

### Table 1 | Major demonstrated DNA-recording approaches

<table>
<thead>
<tr>
<th>Device architecture</th>
<th>Recombinase</th>
<th>ssDNA editing</th>
<th>Base editing</th>
<th>Cas9 array collapse</th>
<th>Self-targeting gRNA</th>
<th>Polymerase ticker tape</th>
<th>CRISPR array</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Detection:</strong> input signals</td>
<td>Transcriptional</td>
<td>Transcriptional</td>
<td>Transcriptional</td>
<td>Transcriptional</td>
<td>Transcriptional</td>
<td>Ion concentration</td>
<td>Transcriptional</td>
</tr>
<tr>
<td><strong>Transformation:</strong></td>
<td>Expression of recombinase</td>
<td>Conversion to ssDNA via retron</td>
<td>Expression of base editor and gRNA</td>
<td>Expression of Cas9</td>
<td>Expression of Cas9</td>
<td>Alteration of misincorporation rate</td>
<td>Conversion to DNA via copy-inducible plasmid</td>
</tr>
<tr>
<td><strong>DNA writing:</strong> bit address</td>
<td>Multiple base pairs</td>
<td>Single or multiple base pairs</td>
<td>Single base pair</td>
<td>Multiple base pairs</td>
<td>Multiple base pairs</td>
<td>Multiple base pairs</td>
<td>Multiple base pairs</td>
</tr>
<tr>
<td><strong>DNA writing:</strong> flexible address</td>
<td>Fixed; flanked by recognition sites</td>
<td>Flexible; ssDNA targeted</td>
<td>Flexible; dCas9 targeted</td>
<td>Fixed; engineered target array</td>
<td>Fixed; self-targeting gRNA sequence</td>
<td>Fixed; polymerase target</td>
<td>Fixed; CRISPR array</td>
</tr>
<tr>
<td><strong>DNA writing:</strong> write operation</td>
<td>Defined; inversion, excision or integration</td>
<td>Defined; ssDNA and/or recombinase-mediated allelic replacement</td>
<td>Defined; C·G-to-T·A base editing</td>
<td>Stochastic; collapse of target array by random indels</td>
<td>Stochastic; sequence evolution of gRNA by random indels</td>
<td>Stochastic; patterns of polymerase misincorporation across target sequence</td>
<td>Directional; integration of CRISPR spacers</td>
</tr>
<tr>
<td><strong>Reading and actuation:</strong></td>
<td>Allele PCR, sequencing and functional response</td>
<td>Sequencing and functional response</td>
<td>Sequencing and functional response</td>
<td>Sequencing</td>
<td>Sequencing</td>
<td>Sequencing</td>
<td>Sequencing</td>
</tr>
</tbody>
</table>

**Performance metrics**

- **In vivo temporal resolution:** +
- **In vivo storage capacity:** ++
- **Demonstrated host range:** Prokaryotic and eukaryotic
- **Representative implementations:**
  - Refs: RSM\(^7\), BLADE\(^7\), Scribe\(^3\), CAMERA\(^5\), GESTALT\(^1\), MEMOIR\(^7\), mSCRIBE\(^4\), Zamft et al.\(^5\), Shipman et al.\(^7\), TRACE\(^3\)

Contemporary recording approaches are classified into seven major approaches, and properties falling under each of the recording device components are delineated. In addition, the demonstrated performance metrics of each device are qualitatively assessed (+, low performance; ++, medium performance; +++ (theoretical), high performance). BLADE, Boolean logic and arithmetic through DNA excision; CAMERA, CRISPR-mediated analog multi-event recording apparatus; dCas9, catalytically dead Cas9; GESTALT, genome editing of synthetic target arrays for lineage tracing; gRNA, guide RNA; indels, small insertions or deletions; MEMOIR, memory by engineered mutagenesis with optical in situ readout; mSCRIBE, mammalian SCRiBE; RSM, recombinase state machine; SCRiBE, synthetic cellular recorder integrating biological events; ssDNA, single-stranded DNA; TRACE, temporal recording in arrays by CRISPR expansion.
**Site-specific recombinase systems**

Systems composed of a recombinase enzyme and flanking target recognition sites around a target sequence. These systems enable inversion, excision or integration of the target sequence on the basis of the orientation of recognition sites.

**Recombinase state machine (RSM)**: A fixed-address writer encompassing a formalized architecture of genetic programs created from combinations of three orthogonal recombinase systems.

Most transcription-based biosensors inherently suffer from a lower temporal resolution owing to slow signal transduction and gene expression processes (>10^2 seconds). By contrast, enzyme-based post-translational sensors can respond to signals much quicker (<10^-2 seconds), which may be necessary to capture transient or fast biological processes. In order to rapidly capture signals into DNA, the activity of recording modules must be directly linked to a signal of interest (for example, through chemically inducible dimerization or post-translational modification). Importantly, DNA polymerization can occur at >500 base pairs per second in vivo, which at least theoretically can match the signal transduction speeds of fast biosensors.

**Writing onto DNA medium**

Natural and engineered DNA targeting and modifying enzymes, which include recombinases, polymerases, integrases, nucleases and multifunctional variants, can be leveraged as writing modules in DNA memory systems. Many new molecular tools to manipulate DNA in cells have emerged, with increased programmability, precision and accuracy. The biochemical characteristics of a DNA writer and its accessory factors (exogenous or from the host) define the ‘recording syntax’ of the system, including the base-pair unit of information storage (‘bit’), the sequence location of DNA writing (‘address’) and type of DNA modification employed (‘write operation’) (BOX 2).

**Fixed-address writers**. Fixed-address writers are targeted to specific biological sequences on the basis of the biochemical properties of the DNA-modifying enzyme and work by treating the orientation or presence/absence of specific target DNA sequences as bits or states. Site-specific recombinase systems, which are widely used in gene expression and knockout applications, enable the inversion, excision or integration of specific target DNA sequences depending on the orientation of flanking recognition sites, thereby enabling manipulation of these DNA bits. For example, 11 pairs of orthogonal recombinase systems were mined from metagenomic databases, allowing the creation of a memory array in which each bit is represented by the presence or absence of specified DNA sequences targeted by each recombinase. This system was capable of storing 1.375 bytes of information in the genome of *Escherichia coli* and was further ported to a commensal gut bacterium, *Bacteroides thetaiotaomicron*, for sensing dietary components in the murine gut. As the recombination event is irreversible, integrase–excisionase pairs or complementary recombinase pairs can be utilized to reset the orientation of target addresses. These orthogonal recombinase systems can further be interleaved and layered to achieve more complex functionalities such as counting, signal amplification and digitization or two-input Boolean logic functions.

The complex set of possible combinatorial recombinase target arrangements was recently formalized for three orthogonal recombinase systems in the recombinase state machine (RSM) framework (FIG. 2a). As the recombination process can be stochastic, layered recombinase systems can be utilized to encode information such as the ordering and duration of inputs within a population through the frequency of different recombination states within the population. Finally, complex recombinase arrangements and circuits can be implemented in mammalian systems, demonstrating the portability of fixed-address writing approaches.

**Flexible-address writers**. Unlike fixed-address writers, which are targeted to predefined sequence locations, flexible-address writers are capable of writing to arbitrarily specified and programmable target locations, yielding precise single or multiple base-pair changes. This specifiable nature of flexible-address writers enables a higher density of data storage and more direct interfacing with host programmes and physiology. One implementation is the synthetic cellular recorder integrating biological events (SCRIBE) system demonstrated in bacteria. In SCRIBE, a single-stranded DNA (ssDNA) is first generated by a retron in response to a biological signal. Then, ssDNA...
Fig. 2 | Examples of DNA-recording devices. The functionality of four exemplary DNA-based recorders are illustrated following the recording device architecture.

a | Recombinase state machine (RSM) fixed-address writer. Orthogonal recombinases are expressed in response to a signal, and they mediate excision or inversion events at a designed recombinase address (filled triangles are unrecombined sites; unfilled triangles are recombined sites). On the basis of the ordering of inputs, different resulting address sequences can be achieved, which are read out by sequencing or which can mediate functional responses by interleaving genetic parts (promoters, reporter genes or terminators) within the recombinase address.

b | CRISPR-mediated analog multi-event recording apparatus (CAMERA) flexible writer. Single-guide RNAs (sgRNAs) are expressed in response to a signal and direct a base editor to mutagenize specific loci within a genomic address, which can be read out by sequencing. ’d’ signifies a catalytically dead Cas9 protein variant with no nuclease activity.

c | Mammalian synthetic cellular recorder integrating biological events (mSCRIBE) stochastic writer.

d | Temporal recording in arrays by CRISPR expansion (TRACE) directional writer. A signal is converted into altered DNA abundance through the use of a copy-number inducible trigger plasmid (pTrig). Short spacers can be incorporated into a genomic array address in a directional manner, either from the trigger sequence or at a constant rate from genomic or plasmid reference sequences. Resulting arrays can be sequenced and the order and source of spacers can be compared with a model of CRISPR expansion to classify the signal input sequence over time.

Synthetic cellular recorder integrating biological events (SCRIBE). A single-stranded DNA (ssDNA)-recombination-based flexible writing approach.

Retron
A bacterial reverse transcriptase system that produces a molecule that is a hybrid of RNA and single-stranded DNA (ssDNA) called multicopy ssDNA (msDNA).

mSCRIBE (mammalian SCRiBE). A Cas9-nuclease-based stochastic writing approach.

CRISPR-mediated analog multi-event recording apparatus (CAMERA), which employs engineered base editors to generate C·G- to T·A mutations that encode information bits at designated DNA addresses with single-nucleotide specificity (Fig. 2b). Base editing is mediated by transcription of both a catalytically dead Cas9 (dCas9) fused to a cytidine deaminase and guide RNAs (gRNAs) that target to the DNA memory address. The presence of edited bases and their frequency across the population encode both digital and analog information (that is, signal identity and intensity). Because the sequence of the resulting edited memory addresses are reproducibly generated, additional layers of editing can occur in a sequential manner to encode temporal information, which enables more complex recording architectures. Even more excitingly, recently demonstrated adenosine base editors that generate A·T- to-G·C mutations can work in the opposite mutational direction to cytosine base editors. In future systems, cytidine and adenosine base editors could be utilized in combination to enable a powerful capability to rewrite DNA addresses repeatedly.

Stochastic writers. Stochastic writers record biological information by continually altering a target DNA sequence in a semi-random manner. By analysing the extent and nature of sequence changes, the intensity of a signal can be inferred. For instance, the programmable site-specific nuclease Cas9 (REFs 42,43,44) can be used to generate a double-strand break at a target DNA address, which is then repaired by endogenous non-homologous end joining (NHEJ) processes, which at a low probability may yield sequence insertions or deletions (indels). The resulting indels are diverse; hence, information is generated at the modified DNA address.

In one class of such stochastic writers, Cas9 is used to target designed DNA addresses consisting of multiple identical target sites (known as arrays or scratchpads) that are stochastically and irreversibly modified during continuous cellular recording (68–70). This approach has been utilized for large-scale recording and lineage reconstruction in entire animals. Beyond recording cell lineage information, these writers could be extended to record analog signals, such as the amount of gene expression over time, by coupling Cas9 expression to a cellular signal of interest. A variety of other nucleases such as CRISPR-associated endonuclease Cpf1 (REF 71), zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) could be used in a similar manner.

A recursive stochastic writing approach can also be used for continuous cellular recording, with the potential advantage that recording is linked to stochastic evolution rather than collapse of a target sequence. In this class, gRNAs that direct a Cas9-based writer to the DNA sequence will undergo continuous mutagenesis, which encodes the magnitude of a biological signal of interest. Such recording devices have been demonstrated in mammalian cells to record inflammation levels in a xenograft model.

Directional writers. In contrast to the above approaches in which DNA addresses have predefined storage capacities and DNA is specifically edited or stochastically altered, directional writers have the ability to create new DNA sequences through addition of nucleotides in a directional manner. As such, these directional writers are well suited for recording temporally changing biological signals. In general, a temporal data recorder (for example, audio recorder) functions by transforming time-varying signals into physical spacing on a substrate (for example, a magnetic tape strip). Similarly, in directional DNA writing, the duration in the time domain is represented by physical distances between recorded data in base pairs.

One such system is a proposed polymerase-based ticker tape, which is an engineered DNA polymerase that writes temporal signals in the form of misincorporated bases as it directionally replicates across a DNA template. The polymerase error rate can be made sensitive to a signal of interest, such as ion concentrations during recording of neuronal activity, thus allowing for temporal encoding of these signals onto DNA memory substrates. Alternatively, CRISPR acquisition systems that catalyse the incorporation of short DNA spacers in a unidirectional manner into expanding CRISPR arrays can be used to record signals. Such systems have been used to
Catalytically dead Cas9 (dCas9). A modified version of Cas9 that lacks endonuclease activity via engineered point mutations. It can be linked to other effector domains for diverse sequence-specific genome engineering applications.

Non-homologous end joining (NHEJ): An endogenous pathway enabling repair of double-strand breaks (DSBs).

Self-targeting gRNA (stgRNA): A single-guide RNA (sgRNA) that is targeted to its own sequence, which enables stochastic sequence evolution over time.

Directional writers: DNA writing relying on directional addition of single or multiple base pairs.

DNA polymerase: A type of enzyme that replicates DNA polymers on the basis of an existing template DNA by serial addition of individual nucleotides.

Temporal recording in arrays by CRISPR expansion (TRACE): A Cas1-Cas2-based CRISPR spacer acquisition system to record biological signals over time.

Fluorescence resonance energy transfer (FRET): A biochemical mechanism of energy transfer between two chromophores that can be utilized for sequence-specific DNA detection applications.

Memory by engineered mutagenesis with optical in situ readout (MEMOIR): A Cas9-nuclease-based stochastic writing approach with spatial readout by single-molecule RNA fluorescence in situ hybridization (smFiSH).

record oligonucleotide sequences that are electroporated into a bacterial population. Because the ordering of incorporated spacers reflects their exposure to the cells, analysis of the resulting arrays across a population of cells allows for reconstruction of exposure ordering. This approach has been further scaled for the recording and storage of a 2.6-kilobyte animated image in the genomes of a bacterial population. We recently described a system, temporal recording in arrays by CRISPR expansion (TRACE), that utilizes CRISPR spacer acquisition to record biological signals by linking a transcriptional signal of interest within a cell to a copy-number-inducible plasmid (Fig. 2d). With this approach, the temporal exposure history over 4 days could be accurately reconstructed, and temporal recordings could be further multiplexed to record three signals across a population of cells. In a conceptually similar manner to these CRISPR integrase approaches, recombinases can also be used to recursively integrate sequences into a genomic array, with the added benefit of larger and more specific sequences that can be incorporated.

Reading from stored data on DNA

The appropriate method for extracting the stored DNA information is dependent on the recording syntax, basepair resolution and throughput needed to decode the data. Oftentimes, the extracted data may need to be further analysed, interpreted or deconvolved using methods-specific in silico reconstruction tools and algorithms to yield the final useful information.

DNA-sequencing-based readers: DNA sequencing is the most direct way to extract information from DNA-based recording devices. Sanger sequencing can provide low-throughput but high-accuracy sequences of ~800 bp. Nucleotide polymorphism frequencies across a population at specific DNA addresses can also be determined from Sanger chromatograms. Alternatively, NGS can determine the sequence of DNA addresses at a much larger scale, and progress in this arena has enabled analysis of many recent recording devices. Short-read sequencing-by-synthesis (from Illumina) can currently provide the highest throughput and read quality, albeit with a maximum read length of ~600 bp. For DNA addresses with longer lengths (for example, large recombinase-targeted loci), long-read sequencing technologies such as single-molecule real-time sequencing (SMRT; from Pacific Biosciences) or nanopore sequencing (from Oxford Nanopore Technologies) are necessary. Although long-read sequencing modalities currently have a relatively lower throughput and lower quality than more mature short-read NGS platforms, portable instruments such as the MinION nanopore sequencer offer exciting real-time readout of DNA data storage.

Molecular and imaging-based readers: For writers with defined addresses, the presence or absence of specific DNA sequences can be directly determined using simple molecular biology tools, such as allele-specific PCR, restriction digestion assays and fluorescence resonance energy transfer (FRET)-based reporters. Alternatively, direct imaging-based techniques enable probing of recorded data from individual cells in their native spatial context. For example, in the memory by engineered mutagenesis with optical in situ readout (MEMOIR) stochastic writer, single-molecule RNA fluorescence in situ hybridization (smFiSH) of edited CRISPR array addresses enables in situ readout of cellular lineage and endogenous gene expression during cellular differentiation. In addition, a number of emerging in situ sequencing approaches, as well as bulk imaging advancements such as expansion microscopy, will support higher resolution spatial readout of a wide range of recording systems.

Data analysis and reconstruction: The scale, complexity and stochastic nature of DNA recording pose new challenges for data analysis and information reconstruction. Quantitative and statistical modelling of the recording performance is essential for mechanistic understanding of the underlying process and failure modes. For instance, in the mammalian SCRIBE stochastic writing system, sequential sequence changes to stgRNAs were analysed by calculating the transition probability between sequence states. Analysis of these data enabled quantitative understanding of key properties of Cas9-mediated DNA editing and the recording process as well as the identification of editing events that led to undesired inactivation of the device.

Modelling essential recording processes can also aid quantitative data reconstruction and information interpretation. In the TRACE directional writing system, a model of CRISPR spacer expansion from either reference or trigger DNA sources was developed and parameterized using control experiments. This model enabled simulation of all possible temporal input states that were then compared with measured data in a classification scheme, which led to accurate predictions of the temporal input signal. Alternatively, parallel DNA-writing systems can be utilized for temporal signal reconstruction. For example, in the MEMOIR stochastic writing system, a model of the recording process suggested that multiple DNA addresses, which are either edited at a constant rate or in response to a signal, can be utilized to reconstruct temporal exposure histories by comparing the resulting writing across these addresses.

Actuation from recorded data: Beyond simply retrieving recorded information from DNA, an important feature of in vivo DNA-based recording is the possibility of transforming recorded data directly into biological responses. Various genetic circuits can be embedded within the architecture of DNA memory, allowing for direct functional responses when data are written and matched to a predefined pattern. For example, promoters and genes of interest can be interleaved within recombinase circuits, allowing for actuation of responses such as expression of multiple fluorescent reporter genes only after the cells are exposed to a specific series of inputs and the target address achieves a specific configuration. A recording device can also directly alter the genotype of a cell upon storage of a specific data set. In the SCRIBE flexible-address
writing system, inactivating mutations (that is, a premature stop codon) in genes of interest were added or removed, resulting in alteration of cellular phenotypes, such as antibiotic resistance, across a cell population. These cellular actuation strategies enable new classes of programmable genetic circuits that can both chronicle biological conditions and respond to them directly by generating heritable DNA changes and not just transient transcriptional responses.

Assessing performance of recording devices

A DNA recorder’s design architecture and biochemical machineries dictate its performance characteristics (that is, temporal resolution, capacity and accuracy of recording) and system capabilities (for example, host portability and multiplexing). Critical and quantitative assessment of different recording modalities is needed to identify their strengths and weaknesses, suitability for a given application and opportunities for further optimization. Here, we outline key performance metrics and assessment criteria to help stratify and evaluate emerging DNA-based recording devices (TABLE 1).

Quantitative performance metrics

Temporal resolution of recording. Different recording architectures can resolve biological signals at different temporal resolutions, which can be quantified in terms of the frequency of input signal per unit time (that is, in hertz). Recording is fundamentally limited by the timescales of sensing machinery, signal transformation and the speed and efficiency of DNA writing. For example, a fixed-address writing system, which must sense a metabolite and respond by expressing a recombinase protein that mediates inversion of a target DNA sequence, has a lower temporal resolution than a polymerase-based directional writing system that directly records ion concentrations close to the rate of DNA polymerization. Importantly, the temporal resolution of DNA writers can be optimized with rational engineering approaches. To match temporal tracking with organismic development, for example, the genome editing of synthetic target arrays for lineage tracing (GESTALT) stochastic writing system employed various engineered Cas9 array configurations that reduced writing efficiency, thus lengthening the timescales of recording.

Capacity and density of information storage. Storage capacity can be quantified in terms of data size in bits per cell. Most systems (for example, defined and stochastic writers) contain a fixed data capacity that is limited by the size of the predefined DNA target address. By contrast, directional writers can increase their storage capacity on the fly as new sequences are written. Together with the recording syntax, the base-pair editing resolution defines the data density or the amount of stored data in bits per base pair. Single-base-pair editing modalities such as Cas9 base-editor flexible-address writers thus offer a higher information storage density. Information can also be distributed across a population to increase storage capacity; for example, for CRISPR integrase directional writers in which individual cells on average contain a small amount of information, a population is required to reconstruct the signal data.

Accuracy and stability of data storage. Accurate data recording and stable data retention over time are crucial for long-term information storage. DNA recorders with higher writing efficiency can, in general, yield more accurate signal reconstructions because data are more efficiently transformed and stored in the DNA. A distinct characteristic of biological recording is the reliance on stochastic DNA writing and the continuous DNA replication and propagation that occur with high, yet still imperfect, fidelity. The origin and location of DNA storage addresses can also affect long-term stability. Different replication systems and sequences may also have different replication fidelity, and recording syntaxes utilizing arrays with higher sequence similarities may have increased levels of recombination that result in loss of data. To improve stability, different error-correction strategies can be used, such as redundant data storage across a population and reconstruction of consensus information in CRISPR integrase-based recordings of image information.

Cross-species portability and cellular burden

A recording system’s enzymatic machinery governs its portability, which is defined as the degree of functionality in diverse hosts. Many DNA-writing modules may depend on specific host factors or processes. For example, stochastic writers rely on Cas9-mediated indels generated by NHEJ repair processes that are prevalent in eukaryotes but rare in prokaryotes. TheSCRIBE system requires expression of a species-specific recombinase to mediate DNA writing in bacteria, and CRISPR integrase-based writing requires an accessory integration host factor (IHF) for spacer integration in E. coli. On the other hand, base-editing DNA writers directly record data by deaminating DNA bases, relying on the highly conserved cellular replication and repair processes found in both eukaryotes and prokaryotes. Indeed, these base-editing systems have been demonstrated in both E. coli and mammalian cells, suggesting high portability of the approach across different hosts.

Recording may also place a burden on native host processes, which can manifest as changes in growth rate, cell physiology or evolutionary stability. Expression of recording machinery may redirect precious cellular resources, whereas the act of DNA writing itself may induce cellular stress responses. In addition, undesired DNA writing, such as Cas9 off-target cleavage or CRISPR integration at non-target sites, may introduce lethal genomic mutations that reduce cell fitness. Finally, the DNA address itself could place an additional burden on the cell to harbour and maintain a larger amount of DNA. These effects may be accentuated over long multigenerational timescales, during which a recording device may acquire inactivating mutations that reduce this burden. For example, characterization of a recombinase-based writer revealed host adaptation to reduce expression of the recombinase, thus inactivating the device. For robust and long-term functionality, the cellular burden of a recording device must be minimized.
**Multiplexing and scalability of biological recording**

Recording devices can be multiplexed, thus enabling simultaneous measurement and comparison of a large number of biological signals. As most recording devices can be modularly linked to transcriptional input signals, various endogenous and engineered transcriptional sensory systems have been linked to recording systems in parallel. If orthogonal recording machinery exists, or if recording can be directed to distinct DNA addresses, multiple channels of recording can be implemented within a single cell. Alternatively, the same recording machinery could be linked to different input sensors in different barcoded cells to allow multiplex data storage across a population, such as in the TRACE system.

Recording systems may be scaled to store different information modalities or link to complementary biological readouts. Constitutive recording at a basal rate (for example, with stochastic writers) enables applications in lineage tracing. The recorded information can be read out in parallel to other readout modalities. For instance, these same approaches can be readily combined with single-cell RNA sequencing (scRNA-seq) methods. In this example, cell type is inferred from the transcriptome, and lineage information is provided by additionally sequencing the DNA address where recording occurs (or RNA transcript expressed from the address) to compare the molecular identity of a cell with its previous lineage.

**Applications of cellular recording**

DNA-based cellular memory systems can be deployed in a variety of useful ways in basic research and applied fields (Fig. 5). Applications in which measurement and tracking of biologically relevant information at locations that are otherwise difficult, if not impossible, to access are particularly well suited for DNA-based recording systems. To implement these systems in contained environments such as individual bioreactors and host-associated microbiomes, or in open settings such as agricultural crops or buildings, different considerations will need to be evaluated and integrated, such as the mode of signal transformation, the spatiotemporal sensitivity and capacity of recording and the stability of data storage.

**Mapping biological processes**

Direct, large-scale and high-resolution cellular recording enables fundamentally new measurements of biological processes that are normally unobtainable. These new data sets will be crucial for improving our understanding of many complex, interconnected and spatiotemporally diverse biological systems and ecologies. In the microbial biosphere where communities can exist at very high density (for example, 10^{11} cells per gram of faecal matter), measuring and tracking every cell is infeasible. Using microbial DNA-based recorders, one could probe and chronicle colonization and gene expression in specific microbial populations within and between hosts (for example, humans, animals or insects) to gain new and greater insights into their ecology and dynamics. Tracking temporal changes of metabolites such as nutrients in these microbiomes could further reveal facets of microbial physiology and metabolic interactions. Furthermore, delineating exposures to phages and mobile DNA using CRISPR-based recorders could be a powerful new approach for analysing horizontal gene transfer processes in different environments in real time.

As DNA recorders can be deployed in single cells and analysed across populations, relative spatial and historical information can be stored in cells of complex tissues and organs during growth, maintenance and ageing. In developmental biology, DNA-based lineage tracing strategies have already enabled the mapping of organisational development at unprecedented scales and resolutions. Extending these approaches to record relevant biological signals will yield new insights in population and developmental biology, potentially down to the single-cell level. For example, DNA-recording approaches have been applied to measure the relationship of cell-state transition processes and lineage in embryonic stem cells. Extensions of such frameworks to the nervous system of complex animals could enable large-scale biological recording and readout of massively complex signalling networks in neurons to probe complex spatiotemporal processes in the brain.

DNA-based recording could also be implemented in emerging cell therapy applications such as chimeric antigen receptor (CAR) T cells to improve actuation in response to complex input signals and track activation history. Beyond measurements of absolute and relative levels of biological signals, DNA-based recorders could also measure variance of these signals across populations, which often govern key community-wide properties such as stochastic gene expression and microbial persistence phenotypes.

**Ubiquitous cellular sentinels**

A wide range of synthetic biology applications exist for cellular sentinels that utilize DNA-based recording systems. Engineering cells in an ecosystem to passively and continuously monitor intracellular and extracellular states and changes (that is, a black box recorder) over large areas and long periods of time constitutes a powerful strategy for ubiquitous sensing and reporting. However, a key limitation of such sentinel cells thus far has been the reliance on colorimetric, fluorescence or luminescence reporter molecules, which require continuously operating detectors that are generally not portable and scalable. DNA-based recorders are poised to substantially affect this arena, creating an entirely new class of environmental sentinel applications. Various recording paradigms could be implemented in engineered organisms — including bacteria, invertebrates (for example, worms), insects (for example, mosquitoes and bees), plants and mammals and their host-associated microbiomes — in both open and contained settings.

To monitor open environments (for example, terrestrial, aquatic or aerial), engineered recorders could track the persistence and levels of pathogen-associated quorum signals, toxic heavy metals and other biotic signatures of interest for various industries to ensure the health of crops, livestock and fisheries. For such open-environment sensing applications, the safety and dissemination of such synthetic recording devices...
Fig. 3 | Applications of DNA-based biological recorders. **a** | Use cases of DNA-based recording (top) as well as applications across research and applied settings (bottom). **b** | Example utility of DNA-based recorders in the gut microbiome. Engineered cellular memory devices could be utilized for non-invasive multiplex temporal recording of important signals such as nutrient status and microbial-derived and host-derived metabolites. In addition, these recorders could mediate functional actions in response to specific signals or profiles of inputs. SCFA, short-chain fatty acid.
must be rigorously assessed and the proper regulatory frameworks must be developed. DNA-based sentinels that can be applied to different surfaces could be used in biosurveillance and forensic applications to monitor the flow of materials (for example, goods or contrabands) and controlled substances (for example, explosives) across the globe. A distinct advantage of cellular fingerprinting and recording strategies over existing inert chemical markers is the ability to track transient or fluctuating changes in environmental conditions (for example, temperature or humidity), which may occur during transportation.

Other settings such as host-associated environments (for example, humans, livestock and insects) are highly relevant application areas for DNA-based sentinels. DNA-based recording approaches have recently been applied to commensal R. thailandensis to record the availability of dietary nutrients such as rhamnose in the gut. This ability to monitor host function and health status using engineered probiotics in the mammalian gut could enable new health-care applications to non-invasively detect and record infections and biomarkers of inflammation levels. Combining these approaches with actuation systems that are directly linked to these memory modules could yield smarter live-cell diagnostic and therapeutic probiotics that are capable of recording and responding to the spatial distribution and dynamics of difficult-to-measure biomarkers and metabolites.

For contained environments such as microbial and mammalian fermentation reactors or bioremediation systems, engineered cellular sensors and recorders could provide real-time monitoring and diagnosis of cell physiology and metabolism to enhance the productivity of cell factories of different chemicals and drugs as well as provenance tracking of valuable or sensitive strains. These active monitoring and recording approaches could be applied to a variety of built environments such as hospitals, airports and schools to examine the spread of contagious and infectious agents. In the future, DNA-based recording devices could interface with silicon-based electronics to interconvert biologically encoded data with digitally stored information. Combined with fast and economical read–write DNA technologies, these approaches could enable direct control and information transfer between biological and electronic systems.

**Box 3 | Key future challenges for DNA-based recording**

**Highly multiplexed recording of diverse input signals**

New engineered sensors and sensing strategies will be required to detect biological signals of interest that cannot currently be sensed or transformed into recordable information. Approaches to scale the number of channels that can be simultaneously recorded in individual cells and increase storage capacity should be explored, such as specifying the recording of different signals to specific target addresses. Synthetic biology circuits could be applied to integrate multiple cellular signals and report on complex cellular phenotypes.

**Recording fast signals**

How can cellular signals be quickly and modularly linked to activate recording into DNA on post-translational timescales? Timescales of native cellular processes (for example, DNA replication) must be taken into account to capture rapidly fluctuating signals. Synthetic circuit modules such as positive feedback loops could be utilized to capture fast or transient inputs.

**Quantifying variability**

Can recording systems be applied to measure variability of a signal across a population? Statistical modelling and machine learning frameworks could be applied to infer the distribution of a signal magnitude or dynamics across cells and improve the accuracy of data reconstruction.

**Stability of recording systems**

Any engineered recording system will place a burden on cell fitness. The resource requirements and cellular impact of systems must be minimized to avoid altering cellular functionality and to improve long-term stability.

**Rapid and low-cost data readout**

For practical or field applications of cellular recording, reading and interpreting raw data from DNA must be possible with minimal resources and in a fast manner. Low-cost and deployable nucleic acid assays or ubiquitous sequencing paradigms such as nanopore sequencing could be utilized to enable practical or field applications of cellular recording.

**Complex in vivo data processing**

New cellular computation paradigms and genetic circuits must be developed to interface with and interpret the results of cellular recording. These new approaches could close the loop between data recording and actuation of cellular responses as a result of complex multi-signal and dynamic input patterns.
Correspondingly, biological processes and corresponding enzymatic machinery with aspects of these features (that is, non-templated polymerases \(^{141,142}\) and terminal deoxynucleotidyl transferases (TdT) \(^{143,144}\)) should be investigated and leveraged for next-generation recording applications. Other strategies not relying directly on the four natural base pairs can also be investigated; for example, unnatural bases could be used to expand the information density and capacity of recording \(^{41}\).

New measurement modalities drive novel scientific understanding of the fascinating behaviours of the natural world. Biological systems span many length and time scales, posing a challenge to traditional direct measurement paradigms that cannot practically be applied to directly measure and record the trillions of cells within developing organisms or environmental microbiomes. DNA-based recording devices offer an exciting new platform to surmount these challenges with a fundamentally different approach. By leveraging the self-replication and large numbers inherent to biological life, these systems could scale rapidly to record signals of previously immeasurable size and resolution, from mapping signal processing networks in the brain to understanding complex ecological niche utilization strategies in densely populated gut microbiomes. Highly optimized recording architectures, novel DNA-writing approaches and continued progress in the scale and ease of sequencing will further drive rapid progress in engineering recording systems that are capable of capturing larger amounts of information and highly multiplex signals. We envision that such DNA memory devices will catalyse a new field of basic research and applied endeavours to understand and probe complex populations or entire organisms.

Published online: 20 September 2018


Acknowledgements
The authors apologize to colleagues whose work could not be cited owing to space limitations. H.H.W. acknowledges funding from the US National Institutes of Health (1R01AI132403-01), the US Office of Naval Research (N00014-17-1-2353, N00014-15-1-2704), the US National Science Foundation (NSF; MCB-1453219) and the Burroughs Wellcome Fund Pathogenesis of Infectious Disease (PATH; 1016691). R.U.S. is supported by a Fannie and John Hertz Foundation Fellowship and an NSF Graduate Research Fellowship (DGE-11-44155).

Author contributions
Both authors contributed to all aspects of the manuscript.

Competing interests
The authors declare no competing interests.

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

Reviewer information
Nature Reviews Genetics thanks T. Fulga, Y. Liu and Y. Michaels for their contribution to the peer review of this work.